MODELIZACIÓN COMPUTACIONAL DE LAS ALTERACIONES METABÓLICAS EN CÁNCER DE MAMA

TESIS DOCTORAL

Lucía Trilla Fuertes

Programa de Doctorado en Biociencias Moleculares

Universidad Autónoma de Madrid

Madrid, 2019

Departamento de Bioquímica

Facultad de Medicina

Universidad Autónoma de Madrid

MODELIZACIÓN COMPUTACIONAL DE LAS ALTERACIONES METABÓLICAS EN CÁNCER DE MAMA

Memoria presentada para optar al grado de Doctor por la licenciada en Biología

LUCÍA TRILLA FUERTES

Dirigida por:

Dr. Juan Ángel Fresno Vara

Dr. Angelo Gámez Pozo

Realizada en el Instituto de Investigación Sanitaria del Hospital Universitario La Paz-IdiPAZ y Biomedica Molecular Medicine S.L.

Hospital Universitario La Paz

Paseo de la Castellana, 261

28046, Madrid

Don Juan Ángel Fresno Vara, Doctor en Química por la Universidad Autónoma de Madrid e Investigador del Sistema Nacional de Salud en el Hospital Universitario La Paz; Don Angelo Gámez Pozo, Doctor en Biología por la Universidad Autónoma de Madrid, investigador en el Hospital Universitario La Paz y Director de Biomedica Molecular Medicine S.L.,

CERTIFICAN que Doña Lucía Trilla Fuertes, Licenciada en Biología por la Universidad Autónoma de Madrid, ha realizado en el laboratorio de Oncogenética Molecular del Instituto de Genética Médica y Molecular del Hospital Universitario La Paz y en Biomedica Molecular Medicine S.L., bajo nuestra dirección, la presente tesis doctoral titulada:

MODELIZACIÓN COMPUTACIONAL DE LAS ALTERACIONES

METABÓLICAS EN CÁNCER DE MAMA.

Y para que conste, firman la presente en Madrid, a 18 de diciembre de 2018.

Dr. Juan Ángel Fresno Vara

Dr. Angelo Gámez Pozo

Visto Bueno del Tutor de Tesis:

Dr. Ramón Díaz Uriarte

Departamento de Bioquímica

Facultad de Medicina, Universidad Autónoma de Madrid

“Siempre llegarás a alguna parte,

si caminas lo suficiente.”

Alicia en el País de las Maravillas

Lewis Carroll

Agradecimientos

A mis directores de tesis, el Dr. Juan Ángel Fresno y el Dr. Angelo Gámez, que me ofrecieron un proyecto de "hacer celulitas", me pusieron a pasar predictores y no entendía nada. Ahora son ellos los que no me entienden a mí.

Al Dr. Enrique Espinosa, porque con su pregunta de "¿y esto para qué vale?" nos ayuda a mantener los pies en la tierra.

A la Dra. Paloma Maín, el Dr. Hilario Navarro y el Dr. Jorge Martín, por hacer las matemáticas mucho más amigables.

A mis compañeros de laboratorio, en especial a Guille, experto en crítica constructiva, y a Rocío, por adoptarme como pollito en los primeros meses.

A Mariana, porque sin todas esas tardes pegándonos con los modelos del metabolismo este trabajo no habría sido posible.

A Irene (Miss Fuentes para los amigos) por mantenerme al día de todos los cotilleos. Esta tesis hubiese sido mucho más aburrida sin ti.

A mis "amiguis" Óscar y Aurora. Ya soy doctora, esto se merece un concierto para celebrarlo.

A mis amigas, mi pequeño aquelarre: Bea, Laura, Paula, Rachel, Carmen, Lorena y Gema. En especial a Gema porque, como dice Juanan de vez en cuando, tú eres la "culpable" de que esté aquí.

A Cristina porque, aunque ahora sea a un país de distancia, llevamos toda una vida siendo amigas.

A Angelo por su paciencia (y lo que te queda). Por favor, no te vengues con más café con sal.

A mis padres, Enrique y Belén, a los que dedico esta tesis, y a mi hermano Diego por su apoyo incondicional siempre.

Gracias a todos por acompañarme en este camino.

RESUMEN

La reprogramación del metabolismo es un proceso característico del cáncer, habiéndose descrito diferencias a nivel metabólico entre subtipos de cáncer de mama. El objetivo de este trabajo es, por un lado, caracterizar la respuesta de líneas celulares de cáncer de mama a dos fármacos con dianas metabólicas (metformina y rapamicina) y, por otro lado, caracterizar tumores de mama a nivel metabólico combinando datos de metabolómica y proteómica con un modelo computacional del metabolismo.

Las líneas celulares de cáncer de mama mostraron una respuesta heterogénea al tratamiento con metformina y rapamicina, provocando en algunos casos un arresto del ciclo celular. Polimorfismos en el gen *SLC22A1* podrían ser la causa de la sensibilidad a metformina mostrada por las células MDAMB468. Además, el análisis proteómico sugiere que los tratamientos provocan alteraciones de procesos como la transcripción. El modelo computacional del metabolismo predice que el tratamiento con metformina produce una disminución en la proliferación y una activación de enzimas relacionadas con estrés oxidativo, que se validó experimentalmente. Se propone además el método de las actividades de los flujos para comparar patrones de flujos entre condiciones.

En cuanto a la caracterización del metabolismo tumoral, las predicciones del modelo computacional del metabolismo son comparables al conocimiento clínico previo, y confirman la naturaleza más proliferativa de los tumores triples negativos y *TN-like*. Asimismo, es posible asociar las actividades de los flujos con el pronóstico.

La estructura funcional, que ya había sido vista en los modelos gráficos probabilísticos basados en expresión génica y proteómica, se mantiene con los datos de metabolómica. El análisis combinado de los datos de expresión génica y metabolómica permitió establecer relaciones entre genes y metabolitos. Combinando las actividades de los flujos con datos de metabolómica se observó coherencia funcional entre metabolitos y la actividad de flujo asociada.

En este trabajo, se emplearon modelos computacionales junto a datos ómicos para caracterizar en líneas celulares la respuesta a fármacos que afectan al metabolismo y las diferencias a nivel metabólico entre tumores de cáncer de mama. Además se propone un nuevo método para comparar patrones de flujo que ha demostrado su utilidad para caracterizar respuesta a fármacos y proponer nuevos factores con valor pronóstico. Finalmente, se creó una interfaz para llevar a cabo los análisis con el modelo computacional del metabolismo sin necesidad de conocimientos de programación.

SUMMARY

Reprogramming of metabolism is a hallmark of cancer. It is described that breast cancer subtypes present differences in metabolic processes, being drugs targeting metabolism good candidates for treatment of this disease. The aim of this work is, on the one hand, the characterization of the response against two drugs targeting metabolism (metformin and rapamycin) and, on the other hand, the characterization of breast tumors at a metabolic level using metabolomics and proteomics data and computational metabolic models.

Breast cancer cell lines showed a heterogeneous response against metformin and rapamycin, causing a cell cycle disruption. Polymorphisms in *SLC22A1* may be the reason of the sensibility of MDAMB468 to metformin. On the other hand, proteomics analyses suggest differences in functional processes, such as transcription, due to the treatments. Moreover, the metabolic computational model predicts a decrease in growth and an activation of enzymes related with oxidative stress (experimentally validated) caused by metformin. A method to compare flux patterns named flux activities was also proposed.

Predictions from computational metabolic models are comparable with previous clinical knowledge, being more proliferative triple negative and *TN-like* tumors. It was also possible to associate flux activities with prognosis.

Strikingly, the functional structure showed in probabilistic graphical models from gene or protein expression data is remained in metabolomics. Combining gene expression and metabolomics data, it was possible to establish relationships between genes and metabolites. Combining flux activities and metabolomics data coherence was showed between metabolites and the associated flux activity.

In this work, computational models, proteomics, metabolomics and gene expression data were employed to characterize response against drugs targeting metabolism in cell lines and metabolic differences between breast tumors. Additionally, a new method to compare flux patterns was proposed and it has demonstrated its utility in the characterization of response and its association with prognosis. Finally, the creation of an interface allows the management of computational metabolic models.

Lastly, an interface was created in order to manage metabolic computational models without the necessity to know programming.

ÍNDICE

RESUMEN 1
SUMMARY 3
ÍNDICE DE TABLAS 9
ÍNDICE DE FIGURAS 11
CLAVE DE ABREVIATURAS 13
INTRODUCCIÓN 19
1. Factores de riesgo 19
2. Histología del cáncer de mama 19
3. Caracterización clínica y molecular del cáncer de mama 20
3.1 Receptores hormonales positivos (ER+) 20
3.2 Her2+ 21
3.3 Triples negativos (TNBC) 21
3.4 *TN-like* 21
4. Tratamiento del cáncer de mama 22
5. Metabolismo tumoral y efecto Warburg 22
6. Fármacos que afectan al metabolismo 23
6.1 Metformina 23
6.2 Rapamicina 24
7. Método de Chou-Talalay para calcular parámetros farmacológicos 24
8. Experimentos de perturbación 25
9. Proteómica 25
10. Metabolómica 26
11. Modelos gráficos probabilísticos 27
12. *Flux Balance Analysis* 27
12.1 Incorporación de datos de expresión: resolución de las *Gene-Protein Reaction rules* y el *E-Flux* 29
13. Aplicaciones del FBA 32
13.1 Estudio de reacciones esenciales en el sistema 32
13.2 *Dynamic FBA* 32
13.3 *Flux Variability Analysis* 33
14. Limitaciones del FBA: El problema de las múltiples soluciones 33
15. Modelos metabólicos previos en cáncer 33
HIPÓTESIS Y OBJETIVOS 39

MATERIAL Y MÉTODOS 43
1. Bases de datos utilizadas 43
1.1 Datos de proteínas provenientes de muestras clínicas 43
1.2 Datos de metabolómica y de expresión génica utilizados para asociar las dos técnicas ómicas e implicaciones en el *Flux Balance Analysis* 43
2. Cultivos celulares y reactivos utilizados 43
3. Ensayos de viabilidad celular 44
4. Construcción de las curvas dosis-respuesta y cálculo de parámetros farmacológicos 44
5. *Array* de *Single Nucleotide Polymorphisms* 45
6. Experimentos de perturbación 45
7. Experimentos de espectrometría de masas y cromatografía líquida 45
8. Identificación de proteínas y cuantificación *label-free* 46
9. Análisis de expresión diferencial en líneas celulares tratadas y sin tratar 47
10. Experimentos de citometría de flujo 47
11. Modelos gráficos probabilísticos y cálculo de la actividad de los nodos 48
12. Construcción de un modelo computacional de metabolismo tumoral 48
13. Introducción de datos de expresión en el modelo del metabolismo 49
14. Validación del modelo metabólico: *dynamic FBA* y estudios de crecimiento celular basados en datos experimentales 50
15. Remuestreo por Monte Carlo 50
16. Cálculo de las actividades de los flujos 51
17. Ensayo de actividad enzimática de la superóxido dismutasa 51
18. Variación debida a la multiplicidad de soluciones 51
19. Construcción de predictores de recaída a distancia usando la actividad de los flujos y los datos de proteómica de tumores de cáncer de mama 52
20. Creación de una interfaz para facilitar la realización del *Flux Balance Analysis* 52
21. Análisis estadístico de los resultados 52
RESULTADOS 57
1. Diseño del estudio de perturbación en líneas celulares de cáncer de mama 57
1.1 Curvas dosis-respuesta y cálculo de parámetros farmacológicos para cada uno de los fármacos que afectan al metabolismo 57
1.2 Genotipado de polimorfismos en las líneas celulares de cáncer de mama 59
1.3 Caracterización de la respuesta a fármacos contra el metabolismo en líneas celulares de cáncer de mama mediante experimentos de perturbación y proteómica 60
1.4 Experimentos de citometría de flujo 61

1.5 Modelos gráficos probabilísticos en líneas celulares 62
1.6 El FBA predice alteraciones en el crecimiento en las células tratadas con metformina 64
1.7 Validación del modelo del metabolismo 64
1.8 Caracterización de las actividades de los flujos en líneas celulares tratadas y sin tratar 65
1.9 Remuestreo por Monte Carlo 66
1.10 El FBA predice una activación de las enzimas relacionadas con estrés oxidativo en células tratadas con MTF 66
1.11 Las mediciones experimentales de la superóxido dismutasa confirman las predicciones hechas por el FBA 69
1.12 Porcentaje de coincidencia entre el valor más frecuente del remuestreo y la primera solución que proporciona el FBA 69
2. Aplicación del *Flux Balance Analysis* a datos de proteómica provenientes de muestras FFPE de tumores de cáncer de mama 70
2.1 Tasa de crecimiento tumoral predicha mediante FBA empleando datos de proteómica de muestras FFPE de tumores de mama 70
2.2 Predictor de recaída a distancia basado en las actividades de los flujos de las muestras tumorales 71
3. Estudio de asociación de datos de metabolómica con los resultados obtenidos en el *Flux Balance Analysis* y con datos de expresión génica en una cohorte de pacientes de cáncer de mama 74
3.1 Análisis basados en los datos de metabolómica 74
3.2 Combinación de datos de expresión génica con datos de metabolómica 77
3.3 Combinación de datos de metabolómica con datos de actividades de los flujos 78
4. Interfaz FLUX para facilitar la realización del *Flux Balance Analysis* 81
DISCUSIÓN 87
1. Experimentos de perturbación 87
1.1 Respuesta de las líneas celulares de cáncer de mama a fármacos que afectan al metabolismo 87
1.2 Genotipado de polimorfismos 88
1.3 Proteómica en líneas celulares de cáncer de mama tratadas y sin tratar 88
1.4 Caracterización de las diferencias a nivel proteico debidas al tratamiento 89
1.5 Estudio del ciclo celular mediante citometría de flujo 90
1.6 Predicciones del crecimiento tumoral mediante el FBA y estudio de las actividades de los flujos en las líneas celulares 90
1.7 Remuestreo por Monte Carlo 91

1.8 Predicción de la activación de enzimas relacionadas con respuesta a estrés oxidativo en células tratadas con MTF y validación experimental mediante la medición de la actividad de la superóxido dismutasa 92
1.9 Predicción de la activación de la óxido nítrico sintasa en MCF7 tratadas con MTF 92
1.10 Resumen de los resultados establecidos para líneas celulares tratadas con MTF 93
1.11 Resumen de los resultados establecidos para líneas celulares tratadas con RP 93
1.12 Limitaciones del estudio 93
1.13 Novedad del estudio 94
1.14 Porcentaje de coincidencia entre el valor más frecuente del remuestreo y la primera solución que proporciona el FBA 94
2. Aplicación del *Flux Balance Analysis* a datos de proteómica provenientes de muestras FFPE de tumores de cáncer de mama 95
2.1 Tasa de crecimiento tumoral para datos de proteómica de muestras FFPE de tumores de cáncer de mama 95
2.2 Predictor de recaída a distancia basado en las actividades de los flujos de las muestras de proteómica 95
2.3 Limitaciones del estudio 96
2.4 Novedad del estudio 96
3. Estudio de asociación de datos de metabolómica con los resultados obtenidos en el *Flux Balance Analysis* y con datos de expresión génica en una cohorte de pacientes de cáncer de mama 97
3.1 Análisis basados en datos de metabolómica 97
3.2 Análisis combinado de metabolitos y genes 98
3.3 Análisis combinado de metabolitos y actividades de flujo 99
3.4 Limitaciones del estudio 99
3.5 Novedad del estudio 100
4. Interfaz FLUX en GUIDE para facilitar la realización del *Flux Balance Analysis* 100
CONCLUSIONES 105
BIBLIOGRAFÍA 111
ANEXO 1: METABOLITOS ASOCIADOS A CADA UNA DE LAS ACTIVIDADES DE FLUJO DE CADA RAMA DE LA RED 125
ANEXO 2: CÓDIGO DE LA APLICACIÓN FLUX 127
ANEXO 3: PUBLICACIONES 131
Artículos que forman parte de esta tesis doctoral 131
Otros artículos 132

ÍNDICE DE TABLAS

Tabla 1: Concentraciones de fármaco empleadas para construir las curvas dosis-respuesta.... 44
Tabla 2: Información acerca de la reacción de biomasa incluida en la Recon2. 49
Tabla 3: Mediciones de viabilidad celular en seis líneas celulares de cáncer de mama tratadas con MTF y RP. 58
Tabla 4: IC_{50} calculada por CompuSyn para cada una de las líneas celulares y de los fármacos. 59
Tabla 5: Funciones mayoritarias de las proteínas con expresión aumentada o disminuida con respecto al control en células tratadas con MTF. 60
Tabla 6: Funciones mayoritarias de las proteínas con expresión aumentada o disminuida con respecto al control en células tratadas con RP. 60
Tabla 7: Modelo de regresión lineal para predicción de respuesta a RP usando las actividades de los nodos funcionales.. 63
Tabla 8: Valores de crecimiento tumoral o biomasa predichos por el modelo del metabolismo 64
Tabla 9: Modelo de regresión lineal para predicción de respuesta a MTF usando las actividades de los flujos. 65
Tabla 10: Modelo de regresión lineal para predicción de respuesta a RP usando las actividades de los flujos.. 66
Tabla 11: Porcentaje de actividad SPODM medido experimentalmente en cada una de las seis líneas celulares de cáncer de mama tratadas con MTF y sin tratar. 69
Tabla 12: Porcentaje de concordancia para cada una de las muestras entre el valor más frecuente de flujo proveniente del remuestreo por Monte Carlo y el valor obtenido mediante el FBA estándar. 70
Tabla 13: Actividades de flujo relacionadas con recaída a distancia en la cohorte de 96 pacientes de cáncer de mama. 72
Tabla 14: Pesos asignados a cada una de las rutas metabólicas para el predictor de recaída. .. 72
Tabla 15: Regresión de Cox comparando el predictor basado en las actividades de los flujos de la beta-alanina, tetrahidrobiopterina y vitamina A. 73
Tabla 16: Pesos asignados a cada actividad de flujo contenida en el predictor para los tumores TNBC. 74
Tabla 17: Análisis multivariante de Cox comparando el predictor basado en las actividades de los flujos de glucolisis y metabolismo del glutamato en TNBC. 74
Tabla 18: Pesos asignados a cada uno de los metabolitos en el predictor. 75

Tabla 19: Modelo de regresión multivariante de Cox comparando el predictor de supervivencia global basado en los datos de metabolitos. 75
Tabla 20: Pesos asignados al predictor compuesto por la actividad del nodo de metabolismo de lípidos. 77
Tabla 21: Análisis multivariante de Cox comparando el predictor basado en la actividad del nodo de metabolismo de lípidos con los datos clínicos. 77
Tabla 22: Descripción de la asociación de los metabolitos con la función de sus correspondientes nodos. 78
Tabla 23: Pesos asignados según el predictor a cada una de las actividades de los flujos. 80
Tabla 24: Modelo de regresión multivariante de Cox comparando el predictor basado en las actividades de los flujos.. 80

ÍNDICE DE FIGURAS

Figura 1: A. Ecuación del efecto medio derivada de la ley de acción de masas en la que se basa el método de Chou-Talalay. 25
Figura 2: Construcción de la matriz *S* a partir de los coeficientes estequiométricos de las reacciones. 28
Figura 3: Bases del FBA..... 29
Figura 4: Ejemplos de posibles GPR y su representación booleana..... 30
Figura 5: Algoritmos para introducir datos de expresión en modelos metabólicos. 30
Figura 6: Algoritmo *E-flux* modificado..... 49
Figura 7: Flujo de trabajo para introducir datos de expresión en el modelo de metabolismo mediante el *E-Flux*. 50
Figura 8: Flujo de trabajo seguido en el estudio de experimentos de perturbación en líneas celulares de cáncer de mama..... 57
Figura 9: A. Curva dosis-respuesta para MTF B. Curva dosis-respuesta para RP. 58
Figura 10: Porcentaje de células medida en cada una de las fases del ciclo celular en células control y células tratadas con MTF o con RP respectivamente. 61
Figura 11: Modelo gráfico probabilístico obtenido a partir de los datos de proteómica de células de cáncer de mama tratadas con MTF o RP y sin tratar. 62
Figura 12: Actividades de los nodos para las líneas celulares tratadas con MTF comparadas con las células control. 63
Figura 13: Actividades de los nodos para las líneas celulares tratadas con RP comparadas con las células control. 63
Figura 14: Número de células medido experimentalmente en cultivos celulares durante 72 horas frente a las predicciones de crecimiento para ese mismo período provenientes del *dynamic FBA.* 65
Figura 15: Distribución de posibles flujos de la catalasa (CATm) en células control y células tratadas con MTF.. 67
Figura 16: Distribución de posibles flujos de la SPODM en células control y células tratadas con MTF..... 68
Figura 17: Distribución de posibles flujos de la NOS2 en MCF7 control y MCF7 tratadas con MTF..... 68
Figura 18: Tasa de crecimiento tumoral predicha por el FBA al introducir los datos de expresión de proteínas para *ER-true*, *TN-like* y TNBC..... 71

Figura 19: Actividades de flujo con diferencias significativas entre subtipos en los datos de proteómica de pacientes con cáncer de mama. 71
Figura 20: Predictor basado en la actividad de los flujos de las rutas de la vitamina A, la tetrahidrobiopterina y la beta alanina en los datos de proteómica de pacientes de cáncer de mama..... 72
Figura 21: Predictor basado en las actividades de los flujos de la ruta de la vitamina A, la beta alanina y la tetrahidrobiopterina por subtipo molecular..... 73
Figura 22: Predictor en tumores TNBC basado en las actividades de los flujos de las rutas de glucolisis y metabolismo del glutamato. 73
Figura 23: Predictor basado en datos de metabolómica. 75
Figura 24: Red de metabolitos proveniente de los datos publicados por Terunuma *et al.* 76
Figura 25: Actividad de los nodos de la red compuesta por metabolitos. 76
Figura 26: Predictor basado en la actividad del nodo del metabolismo de lípidos. 77
Figura 27: A. Red combinada de datos de expresión génica y datos de metabolómica. B. Red combinada de datos de expresión génica y datos de metabolómica caracterizada funcionalmente. 78
Figura 28: Tasa de crecimiento tumoral predicha mediante FBA para los pacientes de esta cohorte. 79
Figura 29: Actividades de los flujos diferenciales entre ER+ y ER-. 79
Figura 30: Predictor basado en las actividades de los flujos del metabolismo de la glutamina y de la alanina y el aspartato. 80
Figura 31: A. Red resultante de combinar los datos de metabolómica con las actividades de los flujos calculadas para cada una de las rutas metabólicas definidas en la Recon2. B. Red de metabolitos y actividades de los flujos dividida por ramas. 81
Figura 32: Aplicación FLUX creada mediante la GUIDE de MATLAB para realizar FBA, FVA y análisis de *knockouts* sin necesidad de utilizar lenguaje de programación. 82

CLAVE DE ABREVIATURAS

AMPK: AMP proteína quinasa.

BIC: Criterio de información bayesiano. De sus siglas en inglés, *Bayesian Information Criterion*.

SLRD: Supervivencia libre de recaída a distancia.

EM: Espectrometría de masas.

ER: Receptor de estrógenos.

ER+: Receptores hormonales positivos.

FBA: Análisis de balance de flujo. De sus siglas en inglés, *Flux Balance Analysis*.

FDR: Tasa de falsos descubrimientos. De sus siglas en inglés, *False Discovery Rate*.

FFPE: Muestras fijadas en formol y embebidas en parafina. De sus siglas en inglés *Formalin-fixed paraffin-embedded*.

FVA: Análisis de variabilidad de flujo. De sus siglas en inglés, *Flux Variability Analysis*.

GPR: De sus siglas en inglés, *Gene-Protein-Reaction rules*.

GUI: Interfaz gráfica de usuario. De sus siglas en inglés *Graphical User Interface*.

Her2: Receptor de crecimiento epidérmico humano 2.

HIF1α: Factor de hipoxia inducible 1 α.

HR: De sus siglas en inglés, *Hazard Ratio*.

IA: Inhibidores de la aromatasa.

IMPaLA: De sus siglas en inglés, *Integrated Molecular Pathway Level Analysis*.

LDHB: Lactato deshidrogenasa B.

MGP: Modelos gráficos probabilísticos.

MPA: De sus siglas en inglés, *Metabolic Phenotypic Analysis*.

MTF: Metformina.

mTOR: De sus siglas en inglés, *mammalian target of rapamycin*.

m/z: Relación masa/carga.

NOS2: Reacción óxido nítrico sintasa 2.

PL: Programación lineal.

PR: Receptor de progesterona.

ROS: Especies reactivas de oxígeno, de sus siglas en inglés *reactive oxygen species*.

RP: Rapamicina.

SBML: De sus siglas en inglés, *Systems Biology Markup Language*.

SNP: De sus siglas en inglés, *Single Nucleotide Polymorphism*.

SPODM: Reacción superóxido dismutasa.

TCA: Ciclo de los ácidos tricarboxílicos, de sus siglas en inglés *Tricarboxilic Acid Cycle*.

TNBC: Cáncer de mama triple negativo, de sus siglas en inglés *Triple Negative Breast Cancer*.

VEGF: Factor de crecimiento vascular epitelial.

INTRODUCCIÓN

INTRODUCCIÓN

Con un millón de casos nuevos en el mundo al año, el cáncer de mama es el cáncer más común en mujeres (1). Representa el 29% de los nuevos casos de cáncer diagnosticados y su probabilidad de aparición en mujeres con más de 60 años está aumentando (2). El cáncer de mama es, tanto desde el punto de vista clínico como desde el punto de vista molecular, una enfermedad compleja y heterogénea.

1. Factores de riesgo

Los factores que aumentan el riesgo de padecer cáncer de mama (1) son:

- Edad: La incidencia de cáncer de mama aumenta con la edad, duplicándose aproximadamente cada 10 años hasta la menopausia, cuando la incidencia disminuye drásticamente.
- Localización geográfica: Existen diferencias significativas entre las diferentes zonas geográficas. Las diferencias entre Oriente y Occidente están disminuyendo, pero todavía existe una incidencia hasta cinco veces mayor en Europa y Norte América.
- Factores hormonales: Menarquía temprana, menopausia tardía o mayor edad al primer parto son factores que incrementan el riesgo de padecer cáncer de mama.
- Antecedentes familiares.
- Enfermedad benigna previa.
- Estilo de vida: dieta, ingesta de alcohol, tabaquismo, etc.
- Tratamientos hormonales: El uso de anticonceptivos orales y de terapias hormonales sustitutivas aumentan ligeramente el riesgo de padecer cáncer de mama en los siguientes 10 años a su ingesta.

2. Histología del cáncer de mama

Los dos tipos histológicos de cáncer de mama más frecuentes son el ductal, actualmente llamado carcinoma invasivo de tipo no especial (3), y el lobulillar, que representan el 75% y el 15% respectivamente de los casos de cáncer de mama en USA (4). El lobulillar deriva de células epiteliales lobulillares (5). El tumor más frecuente (95% de los casos) es el carcinoma, que puede dividirse en:

- *In situ* o no infiltrantes: Son tumores que permanecen localizados dentro de los conductos lácteos o lobulillos de la mama. No invaden tejidos sanos.
- Invasivos o infiltrantes: Son tumores que atraviesan la membrana basal subyacente e invaden otros tejidos.

Por otro lado, cuando existe más de un tumor en la mama, los tumores pueden ser multifocales o multicéntricos. En el caso de los tumores multifocales, todos los tumores se generan a partir de un único tumor original mientras que en los multicéntricos los tumores presentan varios orígenes.

3. Caracterización clínica y molecular del cáncer de mama

Para el diagnóstico de esta enfermedad se utilizan en clínica tres biomarcadores: el receptor de estrógenos (*ESR1*), el receptor de progesterona (*PGR*) y el receptor de crecimiento epidérmico humano 2 (*ERBB2*), denominados en el contexto clínico ER, PR y Her2 respectivamente. En función de estos biomarcadores se establecen tres subtipos: receptores hormonales positivos (ER+ y/o PR+), Her2+ y triples negativos (TNBC). Además, se tienen en cuenta otros factores como tamaño del tumor, afectación ganglionar, grado del tumor, tipo histológico, estatus de proliferación y tasa de crecimiento. (6, 7).

Por otro lado, existe una clasificación basada en perfiles moleculares que subdivide el cáncer de mama en 5 subtipos moleculares: Luminales A (la mayoría ER+, baja proliferación), Luminales B (generalmente ER+, alta proliferación), Her2 enriquecidos (Her2+), Basales (que engloban entre otros a los triples negativos) y Normales (8-10).

3.1 Receptores hormonales positivos (ER+)

ER y PR son receptores nucleares que regulan la expresión de genes específicos. Los estrógenos ejercen su efecto mayoritariamente a través de los receptores ERα y ERβ (11). Los tumores de mama ER+ se caracterizan por la expresión de estos receptores y comprenden entre el 50 y el 80% de los casos de cáncer de mama. La primera terapia dirigida que demostró una mejora en la supervivencia en mujeres con cáncer de mama de tipo ER+ fue el tamoxifeno, un modulador selectivo del ER que se administra en adyuvancia (12).

Otro tipo de tratamiento endocrino administrado a este tipo de pacientes son los inhibidores de la aromatasa (IA), como letrozol y anastrozol (13, 14), que se unen de manera reversible a la aromatasa, enzima encargada de sintetizar los estrógenos. Posteriormente se demostró la superioridad a nivel de tiempo de progresión y respuesta de los IA con respecto al tamoxifeno, por lo que pasaron a considerarse como fármacos de primera línea para el tratamiento del cáncer de mama de tipo ER+ en mujeres postmenopáusicas (15, 16). El último tratamiento endocrino registrado es fulvestrant, un antagonista puro de ER (17-19).

3.2 Her2+

Her2 es un receptor transmembrana tirosina-quinasa, perteneciente a la familia de los receptores del factor de crecimiento epidérmico, que se encuentra sobreexpresado en el 20% de los tumores de mama. Su presencia confiere un fenotipo agresivo con peor pronóstico (20). Su detección se realiza mediante inmunohistoquímica o hibridación fluorescente *in situ* (21).

La señalización de Her2 promueve la proliferación celular a través de la vía Ras-MAPK e inhibe la muerte celular programada modulada por la ruta PI3K/Akt (22).

El uso de trastuzumab (Herceptin®), un anticuerpo monoclonal contra la porción extracelular de Her2, revolucionó el tratamiento de este tipo de cáncer de mama (23). A pesar de su uso establecido, su mecanismo de acción no es completamente conocido, pero se cree que algunos de los mecanismos implicados pueden ser la prevención de la dimerización del receptor Her2, el aumento de la destrucción del receptor por endocitosis y la escisión de su dominio extracelular (24).

3.3 Triples negativos (TNBC)

Los tumores triples negativos (TNBC) se caracterizan por la ausencia de expresión de ER y PR y la ausencia de sobreexpresión de Her2 (25). Dentro de la clasificación molecular de Perou (8), los TNBC se encuentran en su mayoría englobados dentro del grupo de los basales. Recientemente, se han establecido hasta 4 subtipos moleculares dentro de los TNBC, demostrando que son un grupo muy heterogéneo (26).

Los TNBC comprenden aproximadamente el 15% de todos los casos de cáncer de mama. Se caracterizan por un peor pronóstico, asociado con un aumento de las metástasis, una alta tasa de recaída y menor supervivencia que los tumores ER+. Menos del 30% de las mujeres con TNBC metastásico sobreviven 5 años y casi todas fallecen de su enfermedad a pesar de la quimioterapia, que es la base del tratamiento de estos cánceres (27).

Debido a que son muy proliferativos los TNBC responden bien al tratamiento quimioterápico clásico (28, 29). Sin embargo, a diferencia de los otros subtipos, no responden a tratamientos hormonales ni de tipo anti-Her2, ni disponen de ningún tratamiento farmacológico dirigido, más allá de la quimioterapia tradicional.

3.4 TN-like

Recientemente, nuestro grupo ha descrito un nuevo subtipo dentro de los ER+. Estos tumores, que hemos llamado *TN-like*, presentan un perfil de expresión proteico y una prognosis compa-

rable a las presentadas por los tumores TNBC. Aquellos tumores que siguen teniendo características clínicas y moleculares de ER+ los denominamos *ER-true*. Estos subtipos (*ER-true* y *TN-like*) presentan diferencias a nivel de metabolismo (30).

4. Tratamiento del cáncer de mama

En el caso de tumores localizados el tratamiento es curativo, mientras que en tumores ya metastatizados el tratamiento es paliativo. El principal tratamiento en cáncer de mama es la cirugía, que puede ser de dos tipos: una tumorectomía, centrada en eliminar sólo el tejido tumoral, o una mastectomía, consistente en la extirpación de toda o parte de la mama y regiones aledañas. Además, se suele administrar también radioterapia para eliminar el tejido tumoral remanente. En algunos casos se emplean fármacos quimioterápicos después de la cirugía o la radioterapia, lo que se denomina tratamiento adyuvante y cuyo objetivo es destruir las células tumorales que estén dispersas por el organismo.

Como ya se ha mencionado, en el caso de los tumores con receptores hormonales positivos se emplea terapia hormonal. En mujeres premenopáusicas la elección es tamoxifeno durante un período de cinco años. En mujeres posmenopáusicas, en cambio, se emplean IA. En el caso de los tumores Her2+ se administra trastuzumab combinado con quimioterapia (23). La quimioterapia clásica, basada en antraciclinas y taxanos, se administra en el caso de los tumores con receptores hormonales positivos de alto riesgo, en los Her2+ y en los triples negativos.

Sin embargo, cada vez con más frecuencia, se está optando por la administración de los fármacos en neoadyuvancia (antes de la cirugía) debido a las ventajas que proporciona, como pueden ser la posibilidad de realizar una cirugía más conservadora y una pronta medición de la respuesta (31).

5. Metabolismo tumoral y efecto Warburg

Las células tumorales se caracterizan por presentar un metabolismo alterado (32), uno de cuyos procesos más característicos, la glucolisis aerobia, ya fue descrito por Otto Warburg hace casi un siglo (33). Las células normales en presencia de oxígeno producen la mayor parte de su energía mediante el ciclo de los ácidos tricarboxílicos (TCA) y la respiración mitocondrial. Por el contrario, las células tumorales, incluso en presencia de oxígeno, obtienen su energía mediante la glucolisis y la posterior fermentación del piruvato a lactato, y requieren por tanto un mayor aporte de glucosa (33). Warburg propone que la causa de que la célula adopte esta vía menos eficaz es que existen daños en la cadena mitocondrial (34). Sin embargo, posteriormente se ha visto que la mayoría de las células tumorales no presentan daños en la mitocondria

(35). Esta reprogramación del metabolismo es debida a la alteración de las rutas de algunos proto-oncogenes y genes supresores de tumores como la ruta PI3K/Akt (36) o la producción del factor de hipoxia inducible 1 (HIF1), que estimulan la glucolisis y la fermentación del piruvato a lactato frente al TCA (37).

Por el contrario, la entrada de glutamina en el TCA (concretamente en forma de α-cetoglutarato) genera lactato, lo que se conoce como glutaminolisis. El metabolismo de la glutamina sirve para mantener la disponibilidad de aminoácidos no esenciales y para reponer intermediarios del TCA o ciclo de Krebs (anaplerosis) mientras se genera NADH (38). La glutamina es necesaria para la proliferación celular y su metabolismo está regulado por los niveles del oncogén *MYC* (39, 40).

Recientemente, nuestro grupo ha establecido que existen diferencias significativas en diversos procesos celulares como la proliferación o la adhesión según el subtipo de cáncer de mama (30, 41). Entre estos procesos diferenciales se encuentra el metabolismo de la glucosa, que, además, en líneas celulares caracterizadas como TNBC se encuentra aumentado con respecto a líneas celulares ER+. En este trabajo también se caracterizaron diferencias en los niveles de expresión de 19 microARN entre tumores ER+ y TNBC. Asociado al metabolismo de la glucosa se identificó el miR-449a (41).

6. Fármacos que afectan al metabolismo

Las alteraciones del metabolismo descritas en células tumorales mencionadas previamente han llevado a algunas compañías farmacéuticas a considerar estas rutas como fuente de nuevas dianas terapéuticas para el diseño de futuros tratamientos. Algunos de estos fármacos se encuentran en distintas fases de ensayo clínico (42, 43).

6.1 Metformina

La metformina (MTF) es un fármaco de la familia de las biguanidas utilizado tradicionalmente en el tratamiento de la diabetes mellitus tipo 2 (44). Se cree que su mecanismo de acción se basa en la activación de la AMP proteína quinasa (AMPK) y, por tanto, reduce la lipogénesis y la activación de la acetil-coenzima A carboxilasa, enzima encargada de degradar la acetil-coenzima A (45). Además, la MTF inhibe al complejo I mitocondrial por un mecanismo indirecto aún desconocido (46). En células MCF7 de cáncer de mama la MTF es capaz de inhibir el crecimiento y la proliferación celular (47). Actualmente, la MTF se encuentra en ensayos clínicos de fase III para tratar cáncer de mama en adyuvancia tanto en estadios avanzados (NCT01310231) como tempranos (NCT01101438).

6.2 Rapamicina

La rapamicina (RP) o *sirolimus* es un macrólido fungicida aislado por primera vez a partir de *Streptomyces hygroscopicus* hace más de 20 años y el primer inhibidor de mTOR (al que da nombre: *mammalian target of rapamycin*), una proteína de la familia de las quinasas PI3K y un efector de PI3K. La desregulación de la ruta de mTOR juega un papel fundamental en muchas enfermedades, entre ellas el cáncer. La activación de la ruta PI3K/Akt/mTOR es muy común en cáncer de mama, siendo *PIK3CA* el gen mutado con más frecuencia en tumores ER+ (48). RP y derivados (conocidos como "rapálogos") están en diferentes fases de ensayo clínico para el tratamiento de diversos tipos de tumores (49). Además, está aprobado el uso en la clínica del *everolimus*, uno de estos "rapálogos", en pacientes con cáncer de mama avanzado postmenopáusico ER+ y que presentan resistencia a terapia hormonal (50). Estudios previos demuestran que la RP promueve la apoptosis y un bloqueo del ciclo celular en fase G0/G1 en MCF7 (51). mTOR promueve la captación de glucosa y la glucolisis a través de HIF1α, la ruta de las pentosas y la síntesis *de novo* de lípidos (52).

7. Método de Chou-Talalay para calcular parámetros farmacológicos

El método de Chou-Talalay (53, 54) es un método matemático basado en la ley fisicoquímica de acción de masas y que unifica las cuatro ecuaciones principales en biomedicina: la ecuación de Michaelis-Menten, la de Henderson-Hasselbalch y las teorías de unión a ligando de Hill y Scatchard. Este método permite calcular el sinergismo, la potenciación del efecto o el antagonismo de dos o más fármacos y parámetros farmacológicos como la IC_{50}. Los autores de este método sostienen que es posible construir una curva dosis-respuesta a partir de únicamente dos puntos.

Los parámetros de los que consta este método son: la ecuación del efecto medio, derivada de la ley de acción de masas; la pendiente *m* de la recta que se obtiene en la gráfica del efecto medio, el índice de correlación *r*, que nos da una idea de la calidad de los datos, y, en el caso de estudios que impliquen combinaciones de fármacos, el índice de combinación CI. El CI sirve para determinar qué tipo de interacción hay entre combinaciones de fármacos. Un CI = 1 indica un efecto aditivo, un CI>1, sinergia y un CI<1, antagonismo (Figura 1).

A B

$$\frac{f_a}{f_u} = \left(\frac{D}{D}\right)^m \qquad \frac{(D)_1}{(D_x)_1} + \frac{(D)_2}{(D_x)_2} = C$$

Figura 1: A. Ecuación del efecto medio derivada de la ley de acción de masas en la que se basa el método de Chou-Talalay. D= dosis, Dm= dosis con un efecto medio (IC_{50}, EC_{50}, DL_{50}), m= pendiente de la recta, f_a= fracción afectada por a (por ejemplo, el porcentaje de inhibición), f_u= fracción no afectada. B. Ecuación del índice de combinación (CI) para dos fármacos. D_1= dosis de fármaco 1 que en la combinación D_1+D_2 inhibe un x%. D_2= dosis de fármaco 2 que en la combinación D_1+D_2 inhibe un x%. $(Dx)_1$= dosis de fármaco 1 necesaria para producir un x% de inhibición sin ser combinado con otro fármaco. $(Dx)_2$= dosis de fármaco 2 necesaria para inhibir un x% sin combinarse con otro fármaco. Fuente: Chou et al. (55).

El software CompuSyn (Combosyn.Inc) permite el cálculo automatizado de estos parámetros (55).

8. Experimentos de perturbación

En biología, una perturbación consiste tradicionalmente en inhibir o activar una función de una biomolécula mediante la administración de fármacos, un ARN de interferencia o provocando cambios genéticos o epigenéticos (56). Este tipo de experimentos se utilizan para medir cambios en el sistema provocados por la intervención (57).

Se han realizado un gran número de estudios de perturbación en los últimos años y existen numerosas bases de datos públicas que recogen datos de estos experimentos en diferentes organismos, como por ejemplo *Gene Perturbation Atlas* (58) o *Drug/Cell Line Browser* (59).

9. Proteómica

El término proteómica fue acuñado en 1995 y se define como la caracterización a gran escala de todo el proteoma de una línea celular, un tejido o un organismo (60). Se estima que el genoma humano consta de unos 20.000 genes identificados como codificantes de proteínas (61). Sin embargo, el procesamiento alternativo (*splicing*), modificaciones postraduccionales y otros mecanismos celulares llevan a que un gen pueda codificar múltiples isoformas de estas proteínas (62).

En los últimos años se han producido importantes avances en el campo de la proteómica, permitiendo cuantificar del orden de 10.000 proteínas por muestra, lo que equivale prácticamente a todo el proteoma de la célula (63).

La técnica más utilizada para la cuantificación masiva de proteínas es la espectrometría de masas (EM), que se basa en la medición de la relación masa/carga (m/z) de los iones de los péptidos de las proteínas mediante un espectrómetro de masas. La identificación de cada pép-

tido se basa en que cada compuesto tiene un patrón de fragmentación único. El espectrómetro de masas consta de una fuente iónica que ioniza los péptidos, un analizador de masas que se encarga de la dispersión de los iones en función de su relación m/z y un detector que convierte el haz de iones en una señal eléctrica procesable (62). Los datos obtenidos se analizan mediante herramientas bioinformáticas como Mascot y MaxQuant que permiten la identificación y cuantificación de los péptidos obtenidos en los experimentos de EM (64).

10. Metabolómica

La metabolómica es la más reciente de las técnicas ómicas y consiste en la detección y medición de todos los metabolitos de una muestra (65). El término metaboloma se creó en 1998 para definir la totalidad de metabolitos en una muestra biológica (66) y el término metabolómica fue creado en el año 2000 por Fiehn *et al.* (67). La metabolómica es una disciplina que se centra en el estudio holístico de los metabolitos de un sistema biológico, ya sea una célula, un tejido, un órgano o un organismo completo. El término metabolito se refiere a las pequeñas moléculas (normalmente <1.500 Da) producidas o consumidas en las reacciones químicas que tienen lugar en los seres vivos para mantener la vida. El metaboloma del sistema sería la colección de todos estos metabolitos, y proporciona una lectura funcional del estado fisiológico de un organismo, que vendrá determinado por la suma de sus características genéticas, la regulación de la expresión a nivel de transcriptoma y proteoma y las influencias ambientales (68). Los metabolitos representan los productos finales de las complejas redes que conforman los procesos bioquímicos y se encuentran, por tanto, más cerca del fenotipo. Estos metabolitos pueden cambiar como consecuencia de enfermedades, la exposición ambiental o la nutrición. En un entorno clínico, la metabolómica puede proporcionar valiosas herramientas para el diagnóstico y seguimiento de enfermedades complejas como el cáncer (69, 70).

Las técnicas más comunes para la medición de metabolitos son la EM y la resonancia magnética nuclear. La resonancia magnética nuclear posee la ventaja de ser cuantitativa y no necesitar pasos adicionales en la preparación de las muestras. Sin embargo, su sensibilidad aún es baja a pesar de las recientes mejoras. Las técnicas que emplean la EM se basan en las relaciones m/z de cada metabolito o de sus fragmentos. Los recientes avances en la EM permiten medir cuantitativamente miles de metabolitos a partir de cantidades mínimas de material biológico, habiendo disponibles instrumentos comerciales que permiten una sensibilidad en el rango femtomolar (71-73).

11. Modelos gráficos probabilísticos

Los modelos gráficos probabilísticos (MGP), fueron desarrollados en los campos de la Inteligencia Artificial, las Matemáticas y la Economía. Los MGP, compatibles con datos de alta dimensión, consisten en un modelo gráfico no dirigido basado en el criterio de información bayesiano (BIC). La obtención de un MGP se basa en la creación de un árbol de expansión con la probabilidad máxima y posteriormente una búsqueda hacia delante en la que se van añadiendo aristas hasta obtener un modelo óptimo que reduce el BIC lo máximo posible y conserva la capacidad de descomposición del grafo inicial (74, 75).

Este tipo de redes nos permiten, a partir de los datos de expresión de genes o proteínas y sin necesidad de ninguna otra información, construir una red que relacione los genes/proteínas por sus patrones de expresión. Se ha demostrado que estas redes poseen estructura funcional y permiten estudiar diferencias a nivel de procesos biológicos entre grupos de tumores (30, 41, 76, 77).

Con el fin de poder comparar cuantitativamente los datos provenientes de los nodos funcionales de las redes se ha creado una medida llamada actividad de los nodos, consistente en el promedio de la expresión de los componentes (genes, proteínas, etc.) del nodo asociados a la función principal asignada a cada nodo, que ha demostrado su utilidad a la hora de comparar diferentes grupos de tumores. Esta metodología ya ha sido aplicada con éxito a cáncer de mama y a cáncer de vejiga músculo-invasivo (30, 41, 76).

12. *Flux Balance Analysis*

El *Flux Balance Analysis* (FBA) es un método computacional usado para modelar redes metabólicas (78-80). El FBA calcula el flujo de metabolitos a través de la red. Con este método es posible predecir la tasa de crecimiento de un organismo o la tasa de producción de un metabolito concreto. El FBA está basado en la ley de acción de masas y no requiere de parámetros cinéticos, tan sólo de datos estequiométricos de las reacciones implicadas en el modelo (81).

En sus inicios, esta aproximación matemática se empleó en biotecnología para simular el crecimiento de microorganismos como *Escherichia coli* (82). En los últimos años, con la aparición de reconstrucciones cada vez más completas del metabolismo humano, como la Recon2 (83), el FBA se ha aplicado a otras áreas como los glóbulos rojos (84) o el estudio del efecto Warburg en células tumorales (85). Recientemente se ha completado esta reconstrucción del metabolismo humano añadiendo además datos sobre microbiota y algunas enfermedades (86). Existen diversas bases de datos que proporcionan estas reconstrucciones en formato SBML

(*Systems Biology Markup Language*) como KEGG (https://www.kegg.jp/) o BiGG (http:// bigg.ucsd. edu/).

El primer paso del FBA es construir una matriz *S* de dimensiones *m x n* (*m* metabolitos y *n* reacciones). Esta matriz contiene los coeficientes estequiométricos de cada una de las reacciones implicadas en el modelo, tomando un valor positivo los metabolitos que se producen, un valor negativo los que se consumen y cero aquellos que no están implicados en esa reacción. Además, se construye un vector *v* que contiene los flujos de todas las reacciones de la matriz *S*, es decir, la distribución de flujos del modelo (87). Se asume un estado estacionario en el que *Sv=0*, o lo que es lo mismo, no existe acumulación de metabolitos dentro del sistema (Figura 2). A cada reacción se le impone además un flujo mínimo y un flujo máximo posible, $a_i \leq v_i \leq b_i$, siendo *i* cualquiera de las reacciones del modelo. A este modelo pueden añadírsele diferentes tipos de restricciones que definen el espacio en el que se encontrará la solución al problema. Estas restricciones pueden ser fisicoquímicas, espaciales y topológicas, ambientales o reguladoras (87). El intercambio de metabolitos con el exterior se representa mediante reacciones de intercambio o de transporte.

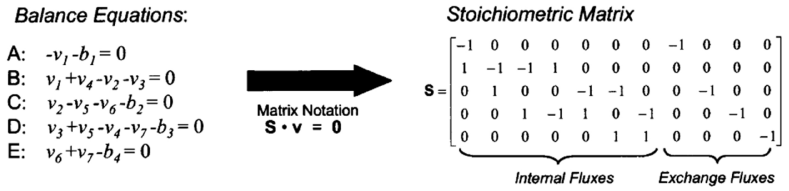


Figura 2: Construcción de la matriz *S* a partir de los coeficientes estequiométricos de las reacciones. Un valor negativo indica que el metabolito se consume en esa reacción, un valor positivo que se produce y un valor 0 que no está implicado en esa reacción. De esta manera se construye la matriz S que incluye tanto reacciones internas como reacciones externas o de transporte. Fuente: Schilling and Palsson (84).

Este problema matemático en concreto se resuelve mediante programación lineal (PL). La PL encuentra una combinación de flujos que maximiza una función objetivo *f* sujeta a restricciones lineales. La función objetivo define el fenotipo de interés para el estudio, como por ejemplo producción de biomasa o tasa de crecimiento tumoral. La estructura del problema sería:

$$\textit{Max } f$$

$$\textit{Teniendo en cuenta que} \begin{cases} Sv=0 \\ a_i \leq v_i \leq b_i \end{cases} \qquad \text{Ec.1}$$

En la red existen dos tipos de reacciones: internas y de intercambio. Las reacciones de intercambio no están equilibradas y representan el suministro o la eliminación de metabolitos por

el sistema al espacio extracelular. Durante la construcción del modelo es necesario especificar cuáles son estas reacciones de intercambio ya que se incorporarán al modelo (87).

La solución obtenida será una combinación lineal de valores de flujo para cada reacción óptima para maximizar la función objetivo. Esta solución estará comprendida entre los límites fijados por las restricciones (81). Existen tres tipos posibles de soluciones: una única solución o sistema compatible determinado, una solución múltiple o sistema compatible indeterminado (diferentes combinaciones de flujos con los que se obtenga el mismo valor óptimo para la función objetivo) y sin solución o sistema incompatible (cuando no es posible encontrar una solución porque la formulación del modelo es incompleta) (Figura 3). El planteamiento y la resolución del problema pueden llevarse a cabo utilizando las diferentes librerías implementadas y de acceso libre como COBRA Toolbox, disponible para MATLAB (88).

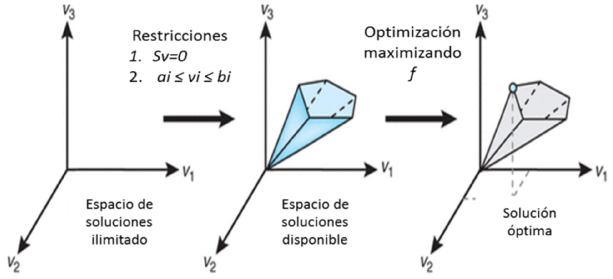
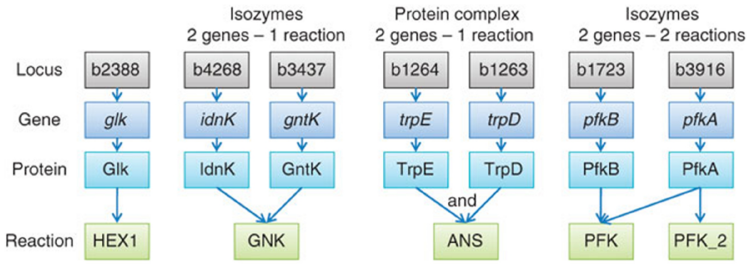


Figura 3: Bases del FBA. En un modelo sin restricciones el espacio no está delimitado por lo que no tiene solución. Al aplicar las restricciones del sistema estacionario Sv=0 y de los límites para cada una de las reacciones $a_i \leq v_i \leq b_i$, siendo i cualquiera de las reacciones del modelo, se acota una región poliédrica entre la que se encuentra la solución al problema de maximizar la función objetivo. Fuente: Orth et al. (81).

12.1 Incorporación de datos de expresión: resolución de las Gene-Protein Reaction rules y el E-Flux

Una de las ventajas del FBA es la posibilidad de incorporar al modelo datos de expresión génica o proteínas. Para ello, es necesario en primer lugar procesar las *Gene-Protein-Reaction Rules* (GPR) incluidas en la Recon2 y que establecen la relación existente entre los diferentes genes implicados en cada una de las reacciones del modelo. Las GPR están formadas por expresiones booleanas compuestas por los operadores *AND* y *OR*. De esta manera, si una reacción es catalizada por isoenzimas (dos enzimas diferentes catalizan la misma reacción) la GPR contendrá un *OR*. Sin embargo, cuando una reacción está catalizada por una proteína con múltiples subunidades sintetizadas por diferentes genes, la GPR contendrá un *AND* (Figura 4).



Reaction abbreviation	Reaction name	E. C.number	GPR
HEX1	Hexokinase (D-glucose:ATP)	2.7.1.1	(b2388)
GNK	Gluconokinase	2.7.1.12	(b3437) or (b4268)
ANS	Anthranilate synthase	4.1.3.27	(b1264) and (b1263)
PFK	Phosphofructokinase	2.7.1.11	(b1723) or (b3916)
PFK_2	Phosphofructokinase (2)	2.7.1.11	(b3916)

Figura 4: Ejemplos de posibles GPR y su representación booleana. La relación puede ser directa (un gen sintetiza una proteína que cataliza una reacción) como en la HEX1. Puede ocurrir que dos isoenzimas catalicen la misma reacción como en el caso de la GNK, siendo el operador booleano *OR*. La reacción puede estar catalizada por un complejo proteico que necesite de dos genes para que sinteticen cada una de las subunidades, en cuyo caso el operador será *AND*, como ocurre en ANS. Por último, puede ocurrir que un gen esté implicado en dos reacciones distintas como es el caso de PFK y PFK2. Fuente: Thiele et al. (89).

En anteriores trabajos, nuestro grupo ha diseñado un método eficiente desde el punto de vista computacional para resolver las GPR (90).

Una vez resueltas las GPR, el siguiente paso consiste en introducir estos datos de expresión ya condensados a nivel de reacción en el modelo. Existen numerosos algoritmos desarrollados para esta tarea pero no existe consenso sobre cuál es el método óptimo. Estos métodos difieren en la forma de procesar los datos: en primer lugar, discretizando los datos de expresión o utilizando el valor continuo y, en segundo lugar, usando valores absolutos para cada condición o relativizando los valores de expresión entre las diferentes condiciones (91) (Figura 5).

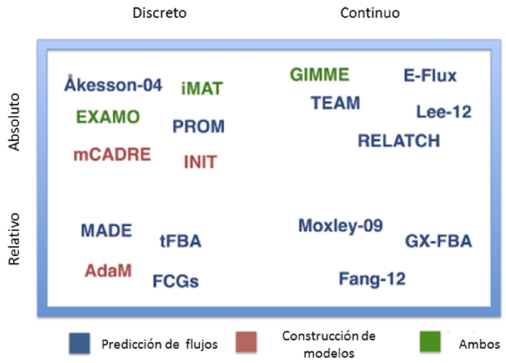


Figura 5: Algoritmos para introducir datos de expresión en modelos metabólicos. Fuente: Machado et al. (91).

Algunos de ellos basados únicamente en datos de expresión (sin tener en cuenta datos de factores de transcripción o cinéticos) son:

- Akesson-04 (2004): Akesson *et al.* proponen limitar el flujo a cero de aquellas reacciones cuyos genes asociados tengan una baja expresión (92).
- GIMME (2008): Se basa en dos pasos: primero se encuentra la distribución de flujos que maximiza la función objetivo y posteriormente se favorece la utilización de aquellas reacciones cuyos genes tengan una expresión por encima de un punto de corte previamente establecido (93).
- iMAT: Este método se basa en la discretización de los valores de expresión génica en tres niveles (bajo, moderado y alto) de acuerdo a un punto de corte establecido previamente por el usuario. iMAT favorece la distribución de flujos que maximice la función objetivo y el uso de las reacciones con genes con alta expresión, minimizando el uso de reacciones que corresponden a genes con bajos niveles de expresión (94).
- *E-flux* (2009): La expresión de cada gen se normaliza dividiendo por el valor máximo de expresión (95, 96). Este algoritmo se explicará más en detalle a continuación.
- Lee-12 (2012): Lee *et al.* proponen integrar los datos de expresión génica como valores absolutos directamente en la función objetivo (97).
- *Metabolic Phenotypic Analysis* (MPA): Método propuesto para introducir datos de proteómica en los modelos metabólicos basado en la discretización de los valores de expresión en bajo, medio y alto (98).

Machado *et al.* realizan una comparativa entre los diferentes métodos empleando datos de transcriptómica provenientes de *Escherichia coli* y *Saccharomyces cerevisiae*, comparándolos con mediciones experimentales. No obtienen resultados concluyentes sobre la superioridad de un método frente a otro. Sin embargo, el *E-flux* se ajusta en múltiples casos a los valores esperados. También comparan la variación en las predicciones al usar datos de expresión génica y proteómica, demostrando que los datos de proteómica proporcionan una mayor exactitud en las predicciones (91).

Por otro lado, Song *et al.* proponen un método basado en la minimización de los flujos, es decir, este método se basa en la asunción de que las células utilizarán el menor número de enzimas posible y que la magnitud de los flujos es proporcional a las concentraciones de enzima. Este método recibe el nombre de *E-Fmin*. Los autores realizan una comparativa de este nuevo algoritmo con los algoritmos anteriores y datos experimentales provenientes de *Saccharomyces cerevisiae* y *Escherichia coli*, obteniendo unos buenos resultados con *E-Fmin*, *E-flux* y Lee-12 (99).

En este trabajo se empleó el algoritmo *E-flux* basado en la relativización de los valores de expresión, que ya había sido utilizado con éxito en trabajos previos para estudiar el efecto Warburg (85). El *E-flux* consiste en incorporar los valores de expresión como restricciones del sistema, siendo el valor de expresión el límite de flujo máximo que puede adquirir esa reacción. Esta aproximación se basa en la asunción de que la cantidad de ARNm es indicativa de la cantidad de enzima disponible para llevar a cabo la reacción y, por tanto, el valor de flujo no puede ser nunca mayor de esta cantidad. Los datos de expresión a_j (para la muestra *j*) se normalizan a un intervalo [0,1] dividiendo por el valor máximo, y se construyen dos vectores correspondientes a los límites mínimo y máximo de cada reacción de manera que si la reacción es irreversible el límite inferior será 0 y si es reversible - a_j y el valor máximo en todos los casos será a_j (96).

En anteriores trabajos en los que estudiamos la mejor manera de normalizar los datos, establecimos como método óptimo de normalización el uso de un algoritmo *E-flux* modificado, consistente en aplicar la función Min-max para normalizar los datos de expresión (90).

13. Aplicaciones del FBA

En sus inicios, el método del FBA se utilizó en biotecnología para optimizar procesos microbiológicos (82). En los últimos años se han ampliado sus aplicaciones a otros ámbitos, entre ellos el estudio de las alteraciones metabólicas en cáncer (85).

13.1 Estudio de reacciones esenciales en el sistema

Una de las aplicaciones del FBA es el estudio de las reacciones esenciales del sistema, consistente en simular en el modelo computacional los efectos de diferentes *knockouts* para determinar cuáles son aquellas reacciones cuya variación produce un efecto en la biomasa (100, 101). Este tipo de análisis se ha empleado recientemente para proponer nuevas dianas terapéuticas (102, 103). La idea que subyace bajo esta aproximación, centrándonos en cáncer, es que si el *knockout* de una reacción produce una disminución de la biomasa, esta reacción es una buena candidata a diana terapéutica por sus posibilidades de disminuir el crecimiento tumoral al ser inhibida.

13.2 Dynamic FBA

Es posible modelar estados dinámicos del sistema mediante una variación del FBA llamada *dynamic FBA*, formulada por Varma *et al*. en 1994 (104). Esta aproximación consiste en la asunción de un estado cuasi-estacionario, teniendo en cuenta la cantidad de biomasa inicial y la concentración de nutrientes disponibles en el medio. Se divide el tiempo experimental en intervalos y se estima el flujo óptimo para cada uno de estos intervalos de manera que el re-

sultado (calculado en condiciones estacionarias) será utilizado como punto de partida del siguiente intervalo, asumiendo así un estado semi-estacionario y pudiendo calcular la variación de biomasa y de concentración de nutrientes en el medio con respecto al tiempo. Esta aproximación ha sido utilizada para la validación de las predicciones del modelo en el caso de Resendis-Antonio *et al.* (105).

13.3 Flux Variability Analysis

El *Flux Variability Analysis* (FVA) nos permite identificar el rango entre el que se encuentran las soluciones óptimas alternativas, entendiéndose como soluciones alternativas todas las posibles combinaciones de flujo que dan lugar a un valor óptimo para la función objetivo. Esta función identifica el máximo y el mínimo valor posible de flujo que puede adoptar una reacción sin variar el valor de la función objetivo, que sigue siendo el óptimo, proporcionando así el rango entre el que se puede encontrar el flujo de esa reacción en las posibles soluciones (81).

14. Limitaciones del FBA: El problema de las múltiples soluciones

Una de las grandes limitaciones del FBA es que este análisis proporciona un único óptimo valor de biomasa, pero existen múltiples combinaciones de flujos que pueden dar lugar a este óptimo. Esto se debe a que el sistema de ecuaciones para resolver este problema es compatible indeterminado, es decir, existen más flujos (o incógnitas) que reacciones (ecuaciones). Con el fin de resolver este problema normalmente se emplea una variación de la técnica de remuestreo Monte Carlo, conocida como *hit and run* (106). Esta aproximación consiste en el cálculo de una distribución aleatoria de posibles soluciones dentro del espacio definido por las restricciones. La librería COBRA Toolbox tiene implementada una función, *gpSampler*, que permite realizar este proceso mediante programación en paralelo, que tiene la ventaja de optimizar el tiempo de procesamiento.

15. Modelos metabólicos previos en cáncer

Existen modelos metabólicos del cáncer de carácter general que han demostrado reflejar la complejidad del metabolismo tumoral. Estos modelos generales recogen las rutas con un papel principal en cáncer (105, 107). Resendis-Antonio *et al.* construyen un modelo simplificado que recoge glucolisis, ciclo de Krebs y ruta de las pentosas y proponen una función objetivo que recoge los principales metabolitos implicados en producción de energía, precursores de aminoácidos y nucleótidos e intermediarios de la glucolisis y la biosíntesis de otros compuestos celulares (105). Mediante este modelo, validado experimentalmente usando células HeLa, estudian posibles dianas metabólicas en cáncer de cérvix. Vázquez *et al.* diseñan un modelo esquemático de producción de ATP y estudian la influencia en el rendimiento del sistema al

variar la cantidad de glucosa suministrada (107). Shlomi *et al.* emplean una reconstrucción completa del metabolismo humano para determinar la importancia de la glucosa y la glutamina en el efecto Warburg (108). En trabajos posteriores este modelo fue empleado para la predicción de nuevas dianas terapéuticas (109).

Jerby *et al*. proponen un método nuevo para introducir datos de proteómica y lo aplican a 392 muestras clínicas de cáncer de mama. Esta aproximación computacional, el *Metabolic Phenotypic Analysis* (MPA), incorpora los datos de expresión dividiéndolos en alta, media o baja expresión (98).

Por otro lado, Asgari *et al.* incorporan datos de expresión génica de 13 líneas celulares de diferentes tipos de cáncer en la reconstrucción completa del metabolismo humano Recon1 utilizando el método del *E-flux* y los analizan mediante FBA (85). Estos resultados muestran que existen diferencias entre los flujos de células tumorales y células normales y que las diferencias son similares entre las diferentes líneas celulares tumorales.

Además, se han construido modelos específicos de tejido usando datos de líneas celulares y tumores específicos. Estos modelos describen rutas diferenciales entre los tumores con respecto al tejido normal (110, 111).

HIPÓTESIS Y OBJETIVOS

HIPÓTESIS Y OBJETIVOS

La reprogramación del metabolismo es uno de los sellos distintivos del cáncer. Recientemente, nuestro grupo ha demostrado que existen diferencias en diversos procesos celulares entre tumores ER+ y TNBC de cáncer de mama, entre ellos el metabolismo de la glucosa. Es lógico suponer que la respuesta a fármacos que afectan al metabolismo será diferente entre subtipos. Los análisis computacionales y modelos metabólicos pueden aportar información complementaria a los análisis convencionales y servir para caracterizar la respuesta a estos fármacos y proponer nuevas dianas terapéuticas.

Los objetivos de este trabajo son:

- Estudio de la respuesta y los efectos del tratamiento de líneas celulares de cáncer de mama a fármacos que afectan al metabolismo mediante experimentos de perturbación y genotipado de polimorfismos.
- Caracterización de la respuesta a fármacos que afectan al metabolismo en líneas celulares de cáncer de mama mediante modelos gráficos probabilísticos.
- Modelización computacional del metabolismo a partir de datos de expresión génica y proteómica en líneas celulares y tumores.
- Evaluación del potencial predictivo de los diferentes modelos a la respuesta farmacológica de las líneas celulares y de la evolución clínica de los pacientes.
- Análisis integrado de datos de metabolómica y de expresión génica provenientes de una cohorte de pacientes de cáncer de mama.
- Aplicación del FBA a esta cohorte empleando los datos de expresión génica y estudio de la relación de los resultados obtenidos con los datos de metabolómica.
- Creación de una interfaz para llevar a cabo el FBA sin necesidad de poseer conocimientos de programación.

MATERIAL Y MÉTODOS

MATERIAL Y MÉTODOS

1. Bases de datos utilizadas

1.1 Datos de proteínas provenientes de muestras clínicas

Se utilizaron datos empleados en trabajos previos de espectrometría de masas de 96 tumores de mama caracterizados previamente como *ER-true*, *TN-like* y TNBC (41). Las muestras, fijadas en formol y embebidas en parafina (FFPE), provenían de los biobancos del Hospital Doce de Octubre y del Hospital La Paz y fueron revisadas por un patólogo para que incluyeran al menos el 50% de células tumorales. Para este estudio se obtuvo la aprobación de los Comités de Ética de ambos hospitales.

1.2 Datos de metabolómica y de expresión génica utilizados para asociar las dos técnicas ómicas e implicaciones en el Flux Balance Analysis

Para la última parte de los análisis se utilizaron los datos del trabajo de Terunuma *et al.* en el que se analizaban tanto por metabolómica como por expresión génica una cohorte de 62 pacientes de cáncer de mama (112). Los datos de metabolómica contienen información acerca de 536 metabolitos. Se calculó el log2 para los datos y se aplicó como criterio de calidad que existiese medición en al menos el 75% de las muestras. Los valores perdidos se imputaron siguiendo una distribución normal usando el software Perseus (113). Después de aplicar estos filtros de calidad, 237 metabolitos se consideraron para los análisis subsiguientes.

En el caso de los datos de expresión génica provenientes de este mismo trabajo, se escogieron los 2.000 genes más variables, es decir, con mayor desviación estándar, para la construcción de los MGP. Estos datos corresponden a un *array* de Affymetrix y se encuentran disponibles en *Gene Expression Omnibus Database* bajo el identificador GSE37751.

2. Cultivos celulares y reactivos utilizados

Para los experimentos con células, se utilizaron tres líneas celulares de cáncer de mama caracterizadas como ER+: MCF7, T47D y CAMA1, y tres líneas celulares caracterizadas como TNBC: MDAMB231, MDAMB468 y HCC1143. Estas líneas celulares se cultivaron en RPMI-1640 con rojo fenol, complementado con 10% de suero fetal bovino inactivado, 1% de estreptomicina y 2% de glutamina, a 37 °C y al 5% de CO_2 en un incubador. Además, las células fueron monitorizadas y autentificadas por características de crecimiento y morfológicas, testadas para *Mycoplasma* y congeladas, y fueron mantenidas en cultivo durante menos de seis meses en todos los experimentos.

Se emplearon dos fármacos que está demostrado que actúan sobre el metabolismo celular: metformina (MTF, Sigma Aldrich D150959) y rapamicina (RP, Sigma Aldrich R8781).

3. Ensayos de viabilidad celular

Las células se descongelaron y se mantuvieron durante 72 horas en el incubador a 37 °C y 5% de CO_2 para estabilizarlas. Posteriormente, se tripsinizaron y contaron mediante el contador de células automático Cell Countess II (Life Technologies) y se sembraron 5.000 células por pocillo en placas de 96 pocillos por triplicado. Pasadas 24 horas se les añadió el fármaco disuelto en el medio de cultivo a diferentes concentraciones y se incubaron durante 72 horas (Tabla 1).

Curvas dosis-respuesta	Concentración MTF (mM)	Concentración RP (nM)
1	0	0
2	5	156,25
3	10	312,50
4	20	625
5	40	1.250
6	80	2.500
7	160	5.000
8	320	10.000

Tabla 1: Concentraciones de fármaco empleadas para construir las curvas dosis-respuesta.

Para la cuantificación de la viabilidad celular tras la exposición al fármaco se utilizó el kit CellTiter 96 Aqueous One Solution Cell Proliferation Assay (Promega). Tras 72 horas de incubación con el fármaco se añadieron 20 µL de CellTiter a cada pocillo, se incubó durante 1 hora a 37 °C, 5% CO_2 y se midió la absorbancia en un lector de placas (TECAN) a una longitud de onda de 490 nm. Como control negativo se usaron pocillos únicamente con medio y como control positivo células sin tratar. Además, se comprobó que el fármaco no interfería en la medición determinando la absorbancia en un pocillo con el fármaco y sin células. La respuesta se midió como el porcentaje de supervivencia de las células con respecto al control, que se consideró el 100%.

4. Construcción de las curvas dosis-respuesta y cálculo de parámetros farmacológicos

Para la construcción de las curvas dosis-respuesta y el cálculo de la IC_{50} se empleó el software CompuSyn (Combosyn. Inc) que permite calcular los diferentes parámetros farmacológicos mediante el método de Chou-Talalay (55) y GraphPad Prism 6. CompuSyn emplea los datos de absorbancia normalizados a un rango [0-1] y las concentraciones de los fármacos empleadas para calcular la IC_{50}.

5. *Array* de *Single Nucleotide Polymorphisms*

Para el genotipado de polimorfismos, se empleó un TaqMan OpenArray en un QuantStudio 12K Flex Real-Time PCR System (Applied Biosystems®) con un formato de *array* que permite el genotipado simultáneo de 180 *Single Nucleotide Polymorphisms* (SNP) en las principales enzimas y los principales transportadores implicados en el metabolismo de los fármacos más frecuentes (PharmArray®). Se recopiló la información sobre las variantes farmacogenéticas asociadas con la respuesta a MTF y RP en la *Pharmacogenomics Knowledge Base* (PharmGKB; www.pharmgkb.org). Se seleccionaron aquellos polimorfismos relacionados con MTF y RP para su análisis en profundidad. La selección final de SNP para nuestro estudio fue la siguiente: rs2032582, rs1045642, rs3213619 y rs1128503 en el gen *ABCB1*; rs55785340, rs4646438 y rs2740574 en *CYP3A4*; rs776746, rs55965422, rs10264272, rs41303343 y rs41279854 en *CYP3A5*; rs1057868 y rs2868177 en *POR* para RP; y rs55918055, rs36103319, rs34059508, rs628031, rs4646277, rs2282143, rs4646278, rs12208357 en *SLC22A1* y rs316019, rs8177516, rs8177517, rs8177507 y rs8177504 en *SLC22A2* para MTF. Además, se llevaron a cabo análisis moleculares mediante técnicas de secuenciación clásicas para los rs34130495 en *SLC22A1* y rs2740574 en *CYP3A4* debido a que estas pruebas no estaban originariamente incluidas en el *array*.

6. Experimentos de perturbación

Para realizar los experimentos de perturbación se utilizaron concentraciones subóptimas de MTF y RP. Como se explicará más adelante en la sección de resultados, las concentraciones empleadas fueron de 40 mM para la MTF, excepto en las MDAMB468 que se usó una concentración de 20 mM, y de 625 nM para la RP. Se sembraron 500.000 células por pocillo en una placa de 6 pocillos. Veinticuatro horas después se añadió el fármaco correspondiente y se incubaron otras 24 horas. Pasado este tiempo se extrajeron proteínas mediante el kit ISOLATE II RNA/DNA/ Protein Kit (BIOLINE) siguiendo las instrucciones del fabricante. La concentración de proteína se midió mediante MicroBCA Protein Assay Kit (Pierce-Thermo Scientific). Los extractos de proteína (10 µg) se digirieron con tripsina (Promega) (1:50). Posteriormente los péptidos se desalaron mediante puntas *stage* C18 caseras. Después se secaron y resuspendieron en 15 µL de acetonitrilo al 3% y ácido fórmico al 1% para el posterior análisis por espectrometría de masas.

7. Experimentos de espectrometría de masas y cromatografía líquida

El análisis de EM se llevó a cabo en un espectrómetro de masas Q-Exactive acoplado a un nano EasyLC1000 (Thermo Fisher Scientific). La composición del disolvente en los dos canales fue de

0,1% de ácido fórmico para el canal A y de 0,1% de ácido fórmico y 99,9% de acetonitrilo para el canal B. Para cada muestra, se cargaron 3 µL de péptidos en columnas caseras (75 µm × 150 mm) empaquetadas con material C18 de fase reversa (ReproSil-Pur 120 C18-AQ, 1,9 µm, Dr. Maisch GmbH) y eluidos con una tasa de flujo de 300 nl/min en un gradiente del 2% al 35% B en 80 minutos, 47% B en 4 minutos y 98% B en 4 minutos. Las muestras se adquirieron con un orden al azar. El espectrómetro de masas se usó en el modo *data-dependent*, adquiriendo un espectro de exploración completo (300−1700 m/z) a una resolución de 70.000 a 200 m/z después de la acumulación a un valor objetivo de 3.000.000, seguido de una fragmentación de disociación de colisión de alta energía en las 12 señales más intensas del ciclo. El espectro HCD se adquirió a una resolución de 35.000 usando una energía normalizada de colisión de 25 y un tiempo máximo de inyección de 120 ms. El control de ganancia automática se fijó en 50.000 iones. Se estableció la detección del estado de carga y se rechazaron los estados de carga no asignados y únicos. Sólo los precursores con una intensidad por encima de 8.300 se seleccionaron para el EM. Las masas precursoras previamente seleccionadas para las mediciones EM se excluyeron de una selección adicional de 30 s, y la ventana de exclusión se fijó a 10 ppm. Las muestras se adquirieron incorporando masas fijas de calibración interna en m/z 371,1010 y 445,2100 (114).

8. Identificación de proteínas y cuantificación *label-free*

Los datos crudos obtenidos en el análisis EM se procesaron mediante MaxQuant (versión 1.4.1.2) (64), seguido de la identificación de proteínas en Andromeda (115). Cada archivo se mantuvo separado en el diseño experimental para obtener valores cuantitativos individuales. El espectro se buscó contra la base de datos *Swiss-Prot*, seguido de una base de datos FASTA de proteínas reversas y contaminantes comunes (NCBI taxonomy ID9606, fecha 2014-05-06). Se introdujeron como variables modificadas la oxidación de la metionina y la acetilación del extremo N-terminal. La especificidad enzimática se fijó para la tripsina/P permitiendo una longitud mínima de 7 aminoácidos y un máximo de dos escisiones perdidas. La tolerancia de fragmento y de precursores se estableció a 10 ppm y 20 ppm respectivamente para la búsqueda inicial. El máximo ratio de falsos descubrimientos (FDR) se fijó en un 1% para péptidos y en un 5% para proteínas. Se habilitó la cuantificación *label-free* y se aplicó una ventana de 2 minutos para la coincidencia entre carreras. Se seleccionó la opción recuantificar. Para definir la abundancia de proteína se utilizó la intensidad, entendida como la suma de las intensidades de los precursores de todos los péptidos identificados para el grupo de proteínas respectivo.

Mediante el software Perseus (113), los datos se transformaron a log2 y se filtraron utilizando como criterios de calidad la presencia de al menos dos péptidos únicos y expresión detectada en al menos el 75% de las muestras. Los valores perdidos se imputaron a una distribución normal.

9. Análisis de expresión diferencial en líneas celulares tratadas y sin tratar

Se compararon los patrones de expresión de proteínas entre las células control y las células tratadas mediante el cálculo de los deltas (Δ) de los valores de expresión para cada fármaco en cada línea celular. En matemáticas, la letra delta delante de una variable indica un cambio en el valor de dicha variable. Los deltas se calcularon restando a la expresión en las células control la expresión en las células tratadas para cada una de las proteínas. Después, se hicieron análisis de ontología para determinar las funciones diferenciales entre células control y células tratadas. Para ello, seleccionamos aquellas proteínas con un delta de expresión mayor de 1,5 o menor de -1,5. La transformación de los identificadores de proteínas a los de genes se realizó en UniProt (http://www.uniprot.org) y DAVID (116). Los análisis de ontología se hicieron marcando *Homo sapiens* como lista de referencia y seleccionando las categorías GOTERM-FAT, Biocarta y KEGG. Se consideraron significativas aquellas categorías con una $p<0,05$ y una FDR por debajo del 5%.

Las relaciones previamente descritas entre genes o proteínas y los fármacos correspondientes se obtuvieron de la *Comparative Toxicogenomics Database* (http://ctdbase.org/) (117).

10. Experimentos de citometría de flujo

Los experimentos se realizaron por duplicado para cada condición. Se sembraron 5.000 células en cada pocillo en placas de 6 pocillos. 24 horas de incubación después se añadieron los fármacos a la concentración subóptima previamente establecida (40 mM en el caso de la MTF y 625 nM para la RP) y después de 72 horas las células se fijaron en etanol y se marcaron con yoduro de propidio. Las células se analizaron mediante un citómetro FACScan equipado con un láser azul a una longitud de onda de 448 nm. Los datos obtenidos se procesaron con BD CellQuest Pro software, primero filtrando las células por tamaño y complejidad para excluir el *debris* y después excluyendo los dobletes y tripletes mediante FL2-W/FL2-A. Estos experimentos se realizaron en colaboración con el Servicio de Citometría de la Universidad Complutense de Madrid.

11. Modelos gráficos probabilísticos y cálculo de la actividad de los nodos

Para la construcción de los MGP se utilizó R v3.5.1 y el paquete *grapHD* (118). Los MGP se construyeron basándose en los datos de expresión, sin otra información *a priori*, y la correlación como medida de asociación. El manejo de la red obtenida mediante R se realizó en Cytoscape (119). La red se dividió en ramas y se realizaron análisis de ontología para establecer la función mayoritaria de cada rama, definiendo diferentes nodos funcionales en la red. Los análisis de ontología para el caso de las redes compuestas por genes o proteínas se llevaron a cabo en DAVID (116) marcando de nuevo *Homo sapiens* como lista de referencia y las categorías GOTERM-FAT, Biocarta y KEGG para el análisis. Para los análisis de ontología relativos a los datos de metabolómica se utilizó IMPaLA (*Integrated Molecular Pathway Level Analysis*) (http://impala.molgen.mpg.de/) (120). Este tipo de análisis se aplicó a los datos provenientes de los experimentos de perturbación en líneas celulares y a los datos de expresión génica y metabolómica provenientes del trabajo de Terunuma *et al.* (112)

Una vez definida la estructura funcional, en el caso de las redes compuestas por genes, proteínas y metabolitos, la actividad de los nodos se calculó mediante el promedio, de la expresión en el caso de genes y proteínas o cantidad en el caso de metabolitos, de cada nodo relacionados con la función o ruta metabólica mayoritaria definida por ontología. En el caso de los experimentos de perturbación en líneas celulares, se calculó la actividad de los nodos mediante el delta de este promedio entre células control y células tratadas. De esta manera, el valor obtenido representa el cambio entre células control y células tratadas a nivel de cada nodo funcional.

12. Construcción de un modelo computacional de metabolismo tumoral

El FBA es un método empleado para la modelización de redes metabólicas que permite tanto la predicción de la tasa de crecimiento de un microorganismo o tumor, como la tasa de producción de un determinado metabolito. Para llevar a cabo el FBA, se utilizó la librería COBRA Toolbox v2.0 (88), disponible para MATLAB y la reconstrucción del metabolismo humano Recon2 (83). En conjunto esta reconstrucción del metabolismo humano consta de 2.191 genes recogidos en las *Gene-Protein-Reaction Rules* (GPR), que incluyen 5.063 metabolitos relacionados mediante 7.440 reacciones, incluyendo las reacciones de intercambio y la reacción de biomasa usada como función objetivo. Esta función objetivo se estableció a partir de datos experimentales provenientes de cultivos celulares de leucemia y es representativa del crecimiento del tumor (Tabla 2). Consideraremos el valor óptimo asignado a esta reacción de bio-

masa como la tasa de crecimiento tumoral. Las 7.440 reacciones están agrupadas a su vez en 101 subsistemas o rutas metabólicas.

Fórmula	0.505626 ala_L[c] + 0.35926 arg_L[c] + 0.279425 asn_L[c] + 0.352607 asp_L[c] + 20.704451 atp[c] + 0.020401 chsterol[c] + 0.011658 clpn_hs[c] + 0.039036 ctp[c] + 0.046571 cys_L[c] + 0.013183 datp[n] + 0.009442 dctp[n] + 0.009898 dgtp[n] + 0.013091 dttp[n] + 0.275194 g6p[c] + 0.325996 gln_L[c] + 0.385872 glu_L[c] + 0.538891 gly[c] + 0.036117 gtp[c] + 20.650823 h2o[c] + 0.126406 his_L[c] + 0.286078 ile_L[c] + 0.545544 leu_L[c] + 0.592114 lys_L[c] + 0.153018 met_L[c] + 0.023315 pail_hs[c] + 0.154463 pchol_hs[c] + 0.055374 pe_hs[c] + 0.002914 pglyc_hs[c] + 0.259466 phe_L[c] + 0.412484 pro_L[c] + 0.005829 ps_hs[c] + 0.392525 ser_L[c] + 0.017486 sphmyln_hs[c] + 0.31269 thr_L[c] + 0.013306 trp_L[c] + 0.159671 tyr_L[c] + 0.053446 utp[c] + 0.352607 val_L[c] -> 20.650823 adp[c] + 20.650823 h[c] + 20.650823 pi[c]
Ruta	Reacción de intercambio

Tabla 2: Información acerca de la reacción de biomasa incluida en la Recon2.

Se calculó el FBA empleando los datos de proteómica provenientes de experimentos de perturbación en líneas celulares, los datos de proteómica de los 96 tumores de cáncer de mama y los datos de expresión génica tomados del trabajo de Terunuma *et al.* (112).

13. Introducción de datos de expresión en el modelo del metabolismo

Para incorporar datos de expresión al modelo se utilizó el algoritmo CAPRI para la resolución de las GPR y el algoritmo *E-flux* modificado, que se habían demostrado óptimos en trabajos previos (90). El algoritmo CAPRI está basado en una modificación del método descrito por Barker *et al.* (121). Las expresiones booleanas que componen las GPR se resolvieron empleando la suma para los *OR* y el valor mínimo para los *AND*. Finalmente, los datos se normalizaron a un intervalo [0,1], restándole al valor de expresión el valor mínimo y normalizando por el rango (función Max-min) (Figura 6). Este valor se utilizó para definir los límites de flujo máximo y mínimo para cada reacción. Si el valor de expresión normalizado para la reacción *r* en la muestra *j* es a_j, entonces los nuevos límites de la reacción serán 0 y a_j si la reacción es irreversible y $-a_j$ y a_j si la reacción es reversible (Figura 7).

$$z = \frac{x - \min(x)}{\max(x) - \min(x)}$$

Figura 6: Algoritmo *E-flux* modificado. z= valor normalizado a un intervalo [0,1], x= valor de expresión, min(x)= valor de expresión mínimo de la muestra, max(x)= valor de expresión máximo de la muestra.

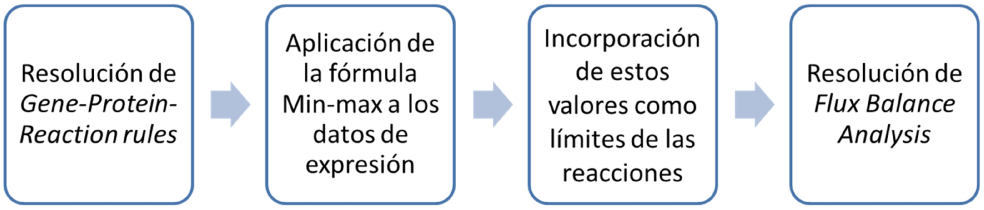


Figura 7: Flujo de trabajo para introducir datos de expresión en el modelo de metabolismo mediante el *E-Flux*.

14. Validación del modelo metabólico: *dynamic FBA* y estudios de crecimiento celular basados en datos experimentales

Para la validación del modelo se llevaron a cabo estudios del crecimiento de las líneas celulares ER+ MCF7 y T47D y de las líneas celulares TNBC MDAMB231 y MDAMB468. Se sembró un millón de células en una placa P100 y se delimitó un área de la placa. A continuación, se cuantificó la densidad celular, es decir, el número de células, de esta área durante tres días (105).

Por otro lado, se empleó el *dynamic FBA* para simular estos cambios en la densidad celular. Para ello se utilizó la función ya implementada *dynamicFBA* a la que es necesario especificar la cantidad de biomasa inicial, correspondiente a la densidad celular inicial en el área de la placa delimitada y la cantidad inicial de glucosa en el medio en el día 1, medida mediante un gasómetro ABL90 FLEX Blood Analyzer, correspondiente a 200 mg/ml.

15. Remuestreo por Monte Carlo

El remuestreo por Monte Carlo se calculó utilizando los datos de líneas celulares tratadas y sin tratar. Para llevar a cabo este remuestreo del espacio de posibles soluciones se utilizó la función *gpSampler* ya implementada en la librería COBRA Toolbox. Esta función permite la realización del proceso de remuestreo en paralelo, optimizando el tiempo de computación. Una vez obtenida la muestra de posibles soluciones, se calculó la suma del valor absoluto de los flujos para cada una de estas soluciones y se escogió aquella con una suma de flujos mayor. Se escogió esta solución como la más representativa debido a la asunción de que si se ha medido una cantidad determinada de una proteína, ésta será usada por la célula y, por lo tanto, la solución de suma de flujos máxima recoge la máxima utilización de las proteínas medidas. Por otro lado, se calculó el *Flux Variability Analysis* (FVA) con la función ya implementada *fluxVariability* y el rango obtenido se utilizó para calcular el cambio en el rango de flujo posible que existía entre células tratadas y sin tratar.

16. Cálculo de las actividades de los flujos

Con el objetivo de comparar la actividad entre las rutas metabólicas en los diferentes escenarios, se calcularon las actividades de los flujos para cada caso. La actividad de los flujos es una medida propuesta en este trabajo, que se definió como la suma de flujos de todas las reacciones implicadas en cada una de las rutas metabólicas definidas en la Recon2.

Con estos datos, en el caso de las líneas celulares, se llevaron a cabo modelos de regresión lineal en SPSS IBM Statistics 20 para asociar estas actividades de los flujos con la respuesta a los fármacos.

En el caso de los datos de proteómica de los 96 tumores, estas actividades de flujo se emplearon para construir predictores de recaída a distancia y en el caso de los datos provenientes del trabajo de Terunuma *et al.* se utilizaron como medida representativa de los resultados del FBA para estudiar su asociación con datos de metabolómica de los mismos pacientes.

17. Ensayo de actividad enzimática de la superóxido dismutasa

Para validar algunas de las predicciones propuestas por el modelo del metabolismo en las líneas celulares se realizó un ensayo de medición de la actividad enzimática de la reacción superóxido dismutasa (SPODM). Estos experimentos se realizaron por triplicado para cada una de las condiciones (cada línea celular tratada con MTF y sin tratar). Para ello se utilizó el Superoxide Dismutase Assay Kit (Sigma-Aldrich, 19160). Se sembraron 5.000 células por pocillo en una placa P6 y, después de 24 horas, se añadió la MTF a 40 mM (excepto para las MDAMB468, en las que se utilizó una concentración de 20 mM, como se explica en resultados). 24 horas después se midió la actividad SPODM siguiendo el protocolo propuesto por el fabricante.

18. Variación debida a la multiplicidad de soluciones

Con el fin de estudiar la variación del vector de flujos sujeto a un valor óptimo de biomasa, se utilizaron los resultados provenientes del remuestreo por Monte Carlo en los experimentos de células tratadas y sin tratar para calcular para cada reacción el valor de flujo más frecuente y si este valor más frecuente difería del valor que aporta la función *optimizeCbModel* (lo que a partir de ahora llamaremos FBA estándar) como primera solución. La función *optimizeCbModel* de COBRA Toolbox está diseñada para dar como solución la primera combinación de flujos que lleva al óptimo que encuentra, que siempre es la misma, haciendo así del FBA un análisis reproducible. Se calculó la moda del flujo para cada reacción en cada uno de los remuestreos y se comparó con el valor obtenido al realizar un FBA estándar en cada uno de los casos. Se cal-

culó el porcentaje de coincidencia C de cada valor, siendo coincidentes (C=1) si cumplían el siguiente criterio: R> (M-0,01) & R< (M+0,01); siendo R el resultado de flujo del FBA y M el valor que toma la moda para ese flujo.

19. Construcción de predictores de recaída a distancia usando la actividad de los flujos y los datos de proteómica de tumores de cáncer de mama

Una vez establecido que la solución aportada por el FBA estándar y las múltiples soluciones obtenidas mediante remuestreo son comparables (ver apartado 1.12 de Resultados), esto hacía posible analizar un gran volumen de datos debido al ahorro de tiempo de computación. Se realizó el FBA con la función *optimizeCbModel* en los datos de proteómica provenientes de 96 tumores de cáncer de mama y se calculó la actividad de los flujos para cada una de las rutas metabólicas definidas en la Recon2. Se estudió la relación de las actividades de los flujos con el riesgo de recaída a distancia y se construyó un predictor con BRB Array Tools (122).

20. Creación de una interfaz para facilitar la realización del *Flux Balance Analysis*

Con el objetivo de que no sean necesarios conocimientos de programación para realizar el FBA, se creó una interfaz gráfica de usuario personalizada (GUI) mediante la GUIDE de MATLAB. GUIDE proporciona herramientas para diseñar interfaces de usuario para Apps personalizadas. Mediante GUIDE es posible diseñar gráficamente la interfaz de usuario y generar de manera automática el código de MATLAB para construir la interfaz (https://es.mathworks.com). Se creó una nueva GUI en la que se incluyeron *Push Buttoms* para cada uno de los pasos necesarios para llevar a cabo el FBA: importar el modelo, importar un archivo en formato *.txt* que contenga las GPR para cada una de las reacciones del modelo, fijar la función objetivo en la reacción de biomasa incluida en la Recon2 y por último calcular el FBA. Además, se le añadieron dos botones para poder realizar el FVA y simulación de *knockouts*. Como control de que el proceso ha terminado se incluyó un texto estático en el que aparece un mensaje de confirmación cada vez que un proceso ha acabado. Por último, se añadió un botón para salir del programa. Para poder usar esta interfaz sólo hay que abrir MATLAB y escribir "FLUX" en la consola. Después, todo el análisis puede llevarse a cabo mediante los botones de la aplicación en lugar de utilizar código.

21. Análisis estadístico de los resultados

Para el análisis estadístico de los datos se utilizó GraphPad Prism 6. Los predictores se construyeron mediante BRB Array Tools, herramienta desarrollada por el grupo del Dr. Richard Simon

(122). Las proporciones de cada uno de los grupos (alto y bajo riesgo) se fijaron *a priori* para evitar un sobreajuste de los predictores a nuestra población. Los modelos de regresión se llevaron a cabo mediante SPSS IBM Statistics 20.

RESULTADOS

RESULTADOS

1. Diseño del estudio de perturbación en líneas celulares de cáncer de mama

Para estudiar el efecto del tratamiento con MTF y RP en seis líneas celulares de cáncer de mama, tres ER+ y tres TNBC, se construyeron curvas dosis-respuesta de cada uno de los fármacos en cada línea celular. Esto permitió definir la concentración del fármaco a emplear en los experimentos de perturbación. Para evaluar si la heterogeneidad en la respuesta a estos fármacos se asociaba a causas genéticas se estudiaron variantes genéticas asociadas a respuesta a estos fármacos mediante la detección de SNP. Para estudiar los procesos biológicos relacionados con la heterogeneidad de respuestas al tratamiento con MTF y RP y caracterizar las consecuencias moleculares de dicha respuesta, se realizaron experimentos de perturbación empleando dosis sub-letales de los fármacos. A continuación, se analizaron los perfiles proteicos mediante EM y se realizó un análisis de expresión diferencial. Con estos datos, se construyó un MGP para caracterizar diferencias debidas a los tratamientos a nivel funcional y se llevó a cabo un FBA para estudiar los efectos de estos tratamientos a nivel de rutas metabólicas (Figura 8). Estas dos aproximaciones (análisis de SNP y proteómica) proporcionan información complementaria acerca de las causas de la respuesta al tratamiento y los efectos moleculares que provocan.

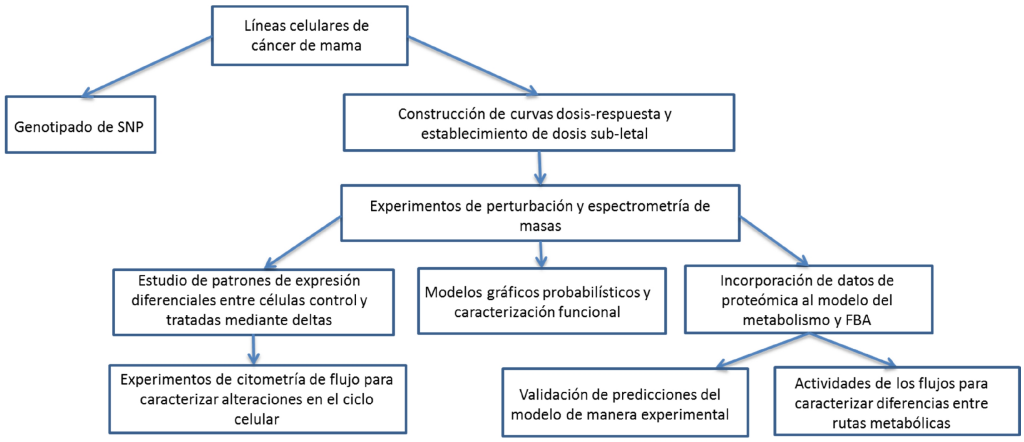


Figura 8: Flujo de trabajo seguido en el estudio de experimentos de perturbación en líneas celulares de cáncer de mama.

1.1 Curvas dosis-respuesta y cálculo de parámetros farmacológicos para cada uno de los fármacos que afectan al metabolismo

Se construyeron curvas dosis-respuesta empleando concentraciones de 5 mM a 160 mM para MTF y concentraciones de 156,25 nM a 10.000 nM para RP en las seis líneas celulares (123). Se midió la respuesta como el porcentaje de células que sobreviven con respecto al control, que se consideró el 100%. Se observó una respuesta diferente para cada una de estas líneas celula-

res que, en el caso de la RP, está asociada al subtipo, teniendo una respuesta menor las líneas celulares TNBC que las ER+. En el caso de la MTF, no se pudo establecer una asociación con el subtipo, siendo las CAMA1 las que presentaban una menor respuesta y las MDAMB468 las más sensibles al tratamiento con MTF (Figura 9, Tabla 3).

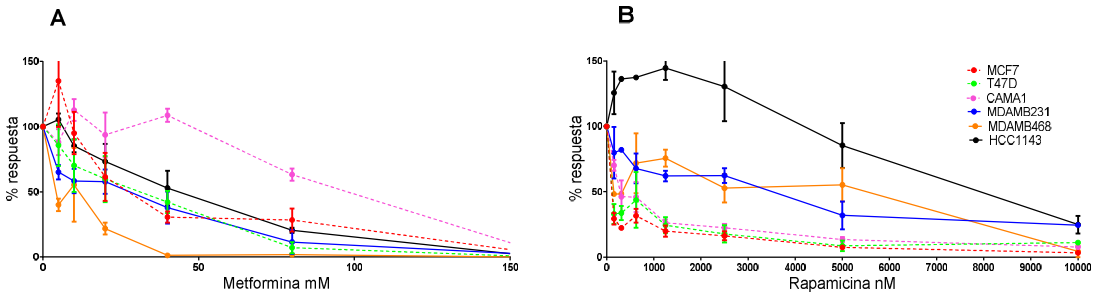


Figura 9: A. Curva dosis-respuesta para MTF B. Curva dosis-respuesta para RP. Las líneas discontinuas corresponden a las líneas celulares ER+ y las continuas a las TNBC.

MTF mM	0	5	10	20	40	80	160
MCF7	100.00	135.07	95.00	61.49	30.61	28.36	2.47
T47D	100.00	85.74	70.15	59.87	42.11	7.10	0.00
CAMA1	100.00	88.08	112.76	93.70	108.67	63.25	3.49
MDAMB231	100.00	65.08	58.36	57.78	37.82	11.45	1.77
MDAMB468	100.00	40.05	55.39	21.82	1.31	1.71	0.00
HCC1143	100.00	105.48	85.25	73.19	52.89	20.49	0.00

RP nM	0	156.25	312.5	625	1250	2500	5000	10000
MCF7	100.00	29.36	22.34	31.62	19.88	16.29	7.53	3.32
T47D	100.00	33.02	33.76	43.74	24.39	17.73	8.69	11.15
CAMA1	100.00	70.22	46.25	45.99	26.28	22.46	13.45	7.71
MDAMB231	100.00	79.92	82.09	67.84	62.16	62.43	31.95	24.50
MDAMB468	100.00	48.25	48.51	71.92	75.75	52.74	55.31	4.49
HCC1143	100.00	125.74	136.39	137.53	144.66	130.58	85.55	24.85

Tabla 3: Mediciones de viabilidad celular en seis líneas celulares de cáncer de mama tratadas con MTF (5-160 mM) y RP (156,25-10.000 nM). Los valores se proporcionan como el porcentaje de células respecto al control (concentración de fármaco =0). Escala rojo-blanco-azul, de mayor a menor.

Las IC_{50} correspondientes a cada línea celular y a cada fármaco se determinaron con CompuSyn (55). Las IC_{50} correspondientes a las células TNBC tratadas con RP eran significativamente más elevadas que las IC_{50} en las células ER+ (Tabla 4).

Células	Subtipo	IC_{50} MTF (mM)	IC_{50} RP(nM)
MCF7	ER+	36,61	62,52
T47D	ER+	17,16	80,43
CAMA1	ER+	87,25	396,573
MDAMB231	TNBC	14,64	2.298,60
MDAMB468	TNBC	6,51	966,32
HCC1143	TNBC	35,19	7.700,24

Tabla 4: IC_{50} calculada por CompuSyn para cada una de las líneas celulares y de los fármacos. En el caso de la RP las líneas celulares ER+ tienen una IC_{50} menor que las TNBC.

Para los experimentos de perturbación se escogieron unas concentraciones que provocaran un efecto en la supervivencia pero no fuesen letales para las células. Se empleó una concentración de 40 mM para la MTF (excepto en el caso de las MDAMB468, que se usó una concentración de 20 mM debido a que son más sensibles a este fármaco) y de 625 nM en el caso de la RP.

1.2 Genotipado de polimorfismos en las líneas celulares de cáncer de mama

Se estudiaron los polimorfismos relacionados previamente con la sensibilidad a MTF y RP mediante un *array* personalizado de sondas TaqMan. El *array* consta de sondas para detectar 180 SNP localizados en 38 genes relacionados con metabolismo y transporte de fármacos. De estos genes, se analizaron aquellos que estaban relacionados con los dos fármacos de interés.

Con respecto a la respuesta a MTF, se detectó el polimorfismo rs2282143 en el transportador *SLC22A1* en las células MDAMB468. Está descrito que este SNP aparece con una frecuencia del 8% en la población negra, que es la población de origen de esta línea celular, y está asociado a una disminución del aclaramiento de la MTF (PharmGKB; www.pharmgkb.org) Por otro lado, el polimorfismo rs628031, también localizado en el gen *SLC22A1*, se encontró en homocigosis en las MCF7 y en las HCC1143 y en heterocigosis con una posible duplicación en las MDAMB468. La presencia de este polimorfismo está asociada con una disminución de la respuesta a MTF (PharmGKB; www.pharmgkb.org).

Respecto a la respuesta a RP, se detectaron dos SNP (rs1045642 y rs2868177) en *ABCB1* y *POR* respectivamente en las líneas celulares ER+. rs1045642 está en heterocigosis en las líneas celulares ER+ y no existe un consenso acerca de sus efectos. Por el contrario, no existe relación descrita de rs2868177 con RP o con otro rapálogo. Las MDAMB468 presentan además un polimorfismo en heterocigosis en *CYP3A4* (rs2740574), que ha sido asociado a un mayor requerimiento de dosis de RP en comparación con el homocigoto *wild-type* (PharmGKB; www.pharmgkb.org).

1.3 Caracterización de la respuesta a fármacos contra el metabolismo en líneas celulares de cáncer de mama mediante experimentos de perturbación y proteómica

Seguidamente se caracterizó molecularmente la respuesta a fármacos contra el metabolismo mediante experimentos de perturbación y proteómica. Se utilizaron seis líneas celulares de cáncer de mama tratadas con concentraciones subóptimas de fármaco (MTF= 40 mM [excepto las MDAMB468, para las que se usó una concentración de 20 mM], RP= 625 nM) y se analizaron por duplicado para cada condición mediante proteómica de alto rendimiento. El análisis por EM permitió la detección de un total de 7.267 proteínas. De éstas, 4.052 proteínas presentaban dos péptidos únicos y expresión en al menos el 75% de las muestras. Estas proteínas se emplearon en los análisis siguientes.

Después, se identificaron aquellas proteínas con una expresión diferencial entre células control y células tratadas. Se seleccionaron aquellas proteínas con un delta en la expresión entre control y tratadas mayor de 1,5 o menor de -1,5 para cada línea celular/fármaco y se realizó un análisis de ontología para establecer en qué funciones estaban implicadas estas proteínas (Tablas 5 y 6).

MTF	MCF7	T47D	CAMA1	MDAMB231	MDAMB468	HCC1143
Disminuida	Mitocondria y ciclo celular	Ninguna	Ninguna	Mitocondria	Mitocondria	Mitocondria y procesamiento de ARNm
Aumentada	Mitocondria y citoesqueleto	Mitocondria y aparato de Golgi	Ninguna	Ninguna	Matriz extracelular	Citosol y unión a proteínas

Tabla 5: Funciones mayoritarias de las proteínas con expresión aumentada o disminuida con respecto al control en células tratadas con MTF.

RP	MCF7	T47D	CAMA1	MDAMB231	MDAMB468	HCC1143
Disminuida	Transporte celular	División celular	Procesamiento de ARNm, *splicing* y mitocondria	Procesamiento de ARNm y citoesqueleto	Ninguna	Lisosomas
Aumentada	Matriz mitocondrial	Lisosomas	Apoptosis, mitocondria y papel mitocondrial en apoptosis	Exosomas	Ninguna	Mitocondria

Tabla 6: Funciones mayoritarias de las proteínas con expresión aumentada o disminuida con respecto al control en células tratadas con RP.

Las proteínas con expresión diferencial entre control y tratamiento se compararon con la información acerca de las interacciones descritas entre estos fármacos y diferentes genes en la base de datos *Comparative Toxicogenomics Database*. De los genes cuya expresión se ve modificada por el tratamiento con MTF en esta base de datos, las proteínas PIR, RELA, SIRT5, CMBL, PPP4R2 y MYD88 presentan una disminución en la expresión, mientras que SRT2, SERPINE1 y

HTATIP2 presentan un aumento en su expresión en células tratadas con MTF en al menos una de nuestras líneas celulares.

Las interacciones recogidas de la *Comparative Toxicogenomics Database* definían una disminución de expresión en los genes que codifican las proteínas CDK4, CKS1B, COL1A1, IGFBP5, KIFC1, mTOR y SCD y un aumento de expresión en CASP8, NR3C1, PKP4, RPS27L, TEAD1 y XIAP debido al tratamiento con RP que observábamos también, a nivel de proteína, en al menos una línea celular tratada con RP.

1.4 Experimentos de citometría de flujo

En los análisis de ontología a partir de las proteínas diferenciales aparecen repetidas veces categorías relacionadas con el ciclo celular, lo que sugiere que estos fármacos provocan de alguna manera alteraciones en dicho ciclo celular. Para confirmar esta hipótesis, se estudió el ciclo celular mediante experimentos de citometría. Las MCF7 y las MDAMB231 tratadas con MTF presentaban un incremento en el porcentaje de células en fase G2/M cuando se comparaban con el control (células sin tratar), sugiriendo que la administración de MTF provoca un arresto del ciclo celular en fase G2. Sin embargo, en las CAMA1, el tratamiento con MTF provocaba un incremento del porcentaje en G0/G1. Con respecto a las células tratadas con RP, las líneas celulares ER+ MCF7 y T47D presentaban un incremento en el porcentaje de células en fase G0/G1 cuando se comparaban con el control, sugiriendo un arresto del ciclo celular en G1 provocado por la RP. Por el contrario, las HCC1143, una línea celular caracterizada como TNBC, presentaba un incremento de células en fase G2 (Figura 10).

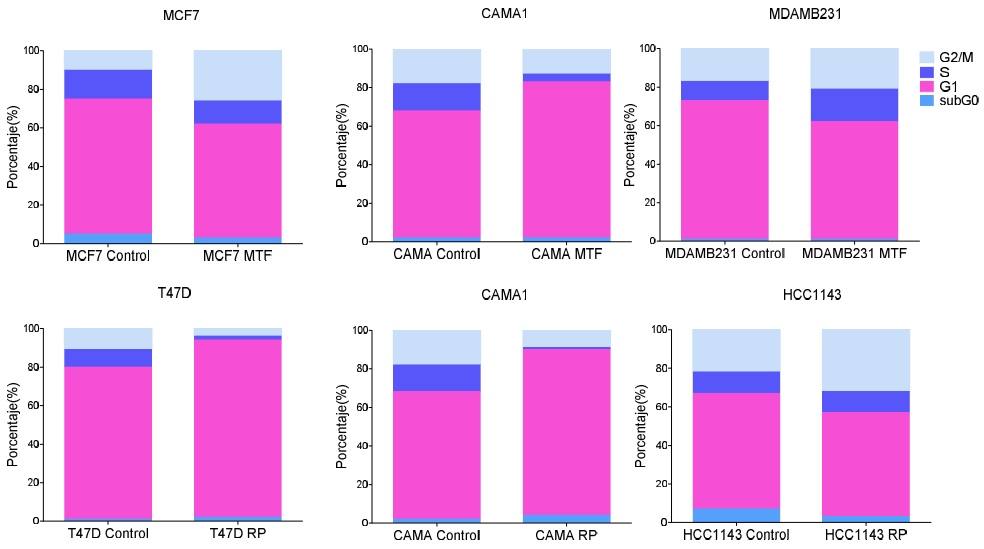


Figura 10: Porcentaje de células medida en cada una de las fases del ciclo celular en células control y células tratadas con MTF o con RP respectivamente.

1.5 Modelos gráficos probabilísticos en líneas celulares

El siguiente paso fue explorar las funciones biológicas afectadas por el tratamiento con MTF y RP. Para ello, se emplearon los datos de proteómica de células tratadas y sin tratar y se aplicaron los MGP sin otra información *a priori*. El grafo resultante se procesó en busca de una estructura funcional, es decir, se establecieron funciones mayoritarias para cada una de las ramas de la red (30, 41, 76). El grafo se dividió en 36 ramas y se realizaron análisis de ontología. Veintinueve ramas tenían un enriquecimiento significativo en alguna función biológica específica mientras que siete ramas carecían de función biológica representativa (Figura 11).

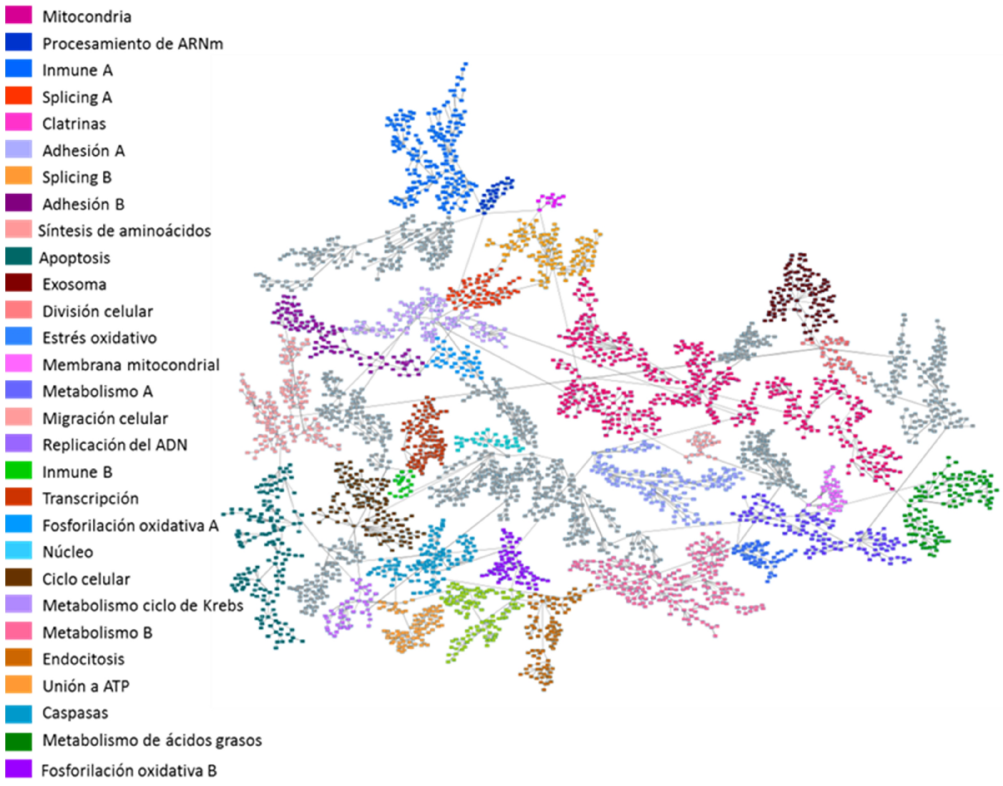


Figura 11: Modelo gráfico probabilístico obtenido a partir de los datos de proteómica de células de cáncer de mama tratadas con MTF o RP y sin tratar. Las ramas a las que no ha sido posible asignarles una ontología aparecen en gris.

Se calcularon las actividades funcionales de cada rama mediante los deltas de la actividad de los nodos entre células control y tratadas. La MTF causaba una disminución en la actividad de los nodos funcionales de mitocondria B, procesamiento de ARNm, replicación del ADN y unión a ATP en todas las líneas celulares (Figura 12). En el caso de la RP, se observó una disminución en la actividad del nodo de procesamiento de ARNm en todas las líneas celulares (Figura 13).

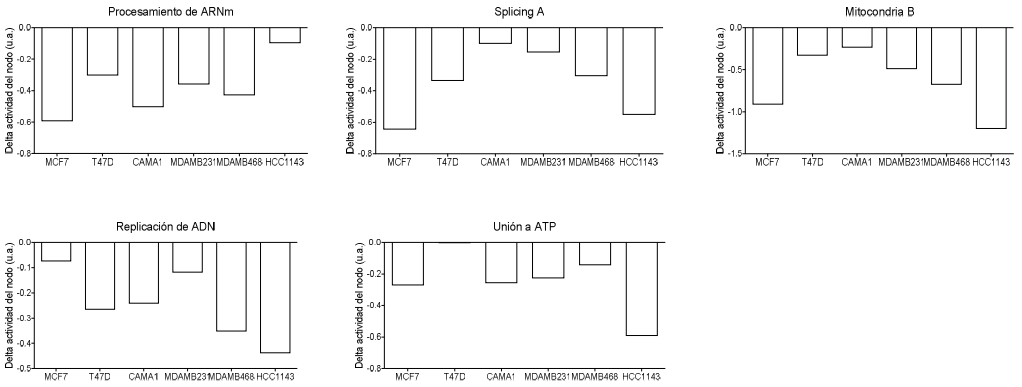


Figura 12: Actividades de los nodos para las líneas celulares tratadas con MTF comparadas con las células control. u.a.= unidades arbitrarias.

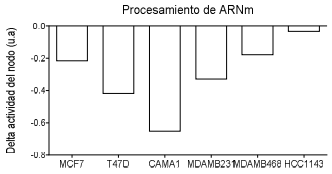


Figura 13: Actividades de los nodos para las líneas celulares tratadas con RP comparadas con las células control. u.a.= unidades arbitrarias.

Además, se evaluó la relación de las actividades de los nodos con la respuesta a MTF y RP mediante modelos de regresión lineal múltiple. La respuesta a RP puede explicarse mediante la actividad de los nodos de metabolismo A y B (Tabla 7). El nodo de metabolismo A está principalmente relacionado con síntesis de ácidos grasos y metabolismo de pirimidinas mientras que el nodo de metabolismo B está principalmente relacionado con glucolisis, fosforilación oxidativa y metabolismo del carbono. La respuesta a MTF no pudo predecirse mediante la actividad de los nodos.

Modelo	B	Error st.	Sig.
Constante	1,095	0,075	0,001
Metabolismo A	-2,171	0,282	0,005
Metabolismo B	-1,149	0,356	0,048

Tabla 7: Modelo de regresión lineal para predicción de respuesta a RP usando las actividades de los nodos funcionales. B = estimación de los coeficientes de regresión, Error st. = error estándar, Sig. = significación estadística al 95% de confianza.

1.6 El FBA predice alteraciones en el crecimiento en las células tratadas con metformina

Para evaluar el impacto en el metabolismo celular tanto de la MTF como de la RP, se realizó un FBA. Para ajustar las predicciones del modelo, se incluyeron los datos de proteómica provenientes de los experimentos de perturbación (datos de expresión de proteínas de las líneas celulares tratadas con ambos fármacos a una concentración fija) y se estimó el crecimiento tumoral tanto para las células control como para las células tratadas. El FBA es un método computacional que permite estudiar la tasa de producción de un metabolito concreto o la tasa de crecimiento de un microorganismo o, en este caso, de un tumor. Los datos de proteómica permitieron delimitar 2.414 reacciones de las 4.253 reacciones contempladas en la Recon2 que tienen definida una GPR. El FBA predice una menor tasa de crecimiento en las MCF7 y en las células TNBC tratadas con MTF con respecto a sus controles, estableciendo el umbral en el segundo decimal. Sin embargo, en el caso de las CAMA1 tratadas con MTF el FBA predice un ligero incremento en la tasa de crecimiento. Por otra parte, el FBA no predice diferencias acordes con las mediciones experimentales en la tasa de crecimiento de las células tratadas con RP con respecto a sus controles (Tabla 8).

Línea celular	Control	MTF	RP
MCF7	0,434	0,425	0,434
CAMA1	0,432	0,438	0,422
MDAMB231	0,453	0,432	0,444
HCC1143	0,440	0,433	0,435

Tabla 8: Valores de crecimiento tumoral o biomasa predichos por el modelo del metabolismo.

1.7 Validación del modelo del metabolismo

Con el propósito de validar el modelo, se realizó una comparación entre datos experimentales de crecimiento en líneas celulares y las predicciones obtenidas mediante el método del *dynamic FBA*. La concentración inicial medida de glucosa en el medio era de 200 mg/dl y fue incorporada a los *inputs* de esta función. La densidad celular inicial se estimó contando las células presentes en un área delimitada de la placa (MCF7= 37, T47D= 31, MDAMB231= 30 y MDAMB468= 58 células) y también se incorporaron como *input* de esta función. Las predicciones hechas por el modelo coinciden con las mediciones experimentales obtenidas durante 72 horas, siendo las MDAMB468 las que más se alejan, mientras que las predicciones para las MCF7 coinciden plenamente con las observaciones experimentales (Figura 14).

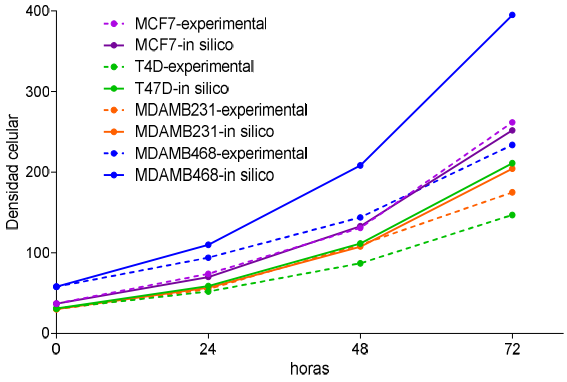


Figura 14: Número de células medido experimentalmente en cultivos celulares durante 72 horas (líneas discontinuas) frente a las predicciones de crecimiento para ese mismo período provenientes del *dynamic FBA* (líneas continuas).

1.8 Caracterización de las actividades de los flujos en líneas celulares tratadas y sin tratar

Con el objetivo de comparar los flujos provenientes de las diferentes rutas metabólicas entre las células tratadas y sin tratar, se estableció un nuevo método llamado actividades de los flujos. Las actividades de los flujos se calcularon como la suma de los flujos de todas las reacciones implicadas en una determinada ruta metabólica, entendiéndose por ruta metabólica aquellas que vienen definidas previamente por la Recon2. De esta manera se resume la información contenida en cada ruta metabólica en un único valor. Una vez calculadas las actividades de los flujos, éstas se utilizaron para construir modelos de regresión lineal con los que predecir respuesta a cada uno de los fármacos.

Para la MTF, las rutas asociadas con respuesta fueron el metabolismo del piruvato y el metabolismo del glutamato (Tabla 9). En el caso de la RP, las rutas metabólicas asociadas con respuesta son el metabolismo del colesterol y la ruta del metabolismo de la valina, leucina e isoleucina (Tabla 10).

Modelo	B	Error st.	Sig.
Constante	0,779	0,002	
Metabolismo de glutamato	-2,379	0,016	0,004
Metabolismo de piruvato	0,083	0,002	0,016

Tabla 9: Modelo de regresión lineal para predicción de respuesta a MTF usando las actividades de los flujos. B = estimación de los coeficientes de regresión, Error st. = error estándar, Sig. = significación estadística al 95% de confianza.

Modelo	B	Error st.	Sig.
Constante	0,761	0,001	
Metabolismo de valina, leucina e isoleucina	2,018	0,002	0,001
Metabolismo de colesterol	0,045	0,000	0,007

Tabla 10: Modelo de regresión lineal para predicción de respuesta a RP usando las actividades de los flujos. B = estimación de los coeficientes de regresión, Error st.= error estándar, Sig. = significación estadística al 95% de confianza.

1.9 Remuestreo por Monte Carlo

El principal problema del FBA es la multiplicidad de soluciones, es decir, para un único valor óptimo de la función objetivo son posibles distintas combinaciones de flujos en las reacciones. Este problema se afronta de diferentes maneras como, por ejemplo, con un remuestreo de un número representativo de posibles soluciones por Monte Carlo y/o escogiendo una combinación de flujos que cumpla una serie de condiciones (por ejemplo, aquella combinación de flujos cuya suma de todos los flujos sea mínima o máxima). Con el propósito de identificar reacciones que se modifican como consecuencia del tratamiento, en este trabajo se realizó un remuestreo por Monte Carlo para cada una de las muestras obteniéndose 14.480 posibles soluciones, y se escogió la solución en la que la suma de flujos fuese el máximo posible. Este criterio se basó en la premisa de que esta combinación sería la más representativa de las mediciones de las proteínas debido a que si una proteína ha sido medida es indicativo de que esta proteína será usada por la célula y por tanto debería estar reflejada en estos flujos. A continuación, se realizó un *Flux Variability Analysis* (FVA) para calcular el valor máximo y mínimo que pueden tomar los flujos de cada una de las reacciones en las múltiples soluciones y, por tanto, el rango de flujos posibles. Después, se eligieron aquellas reacciones que presentaban un cambio en su flujo entre tratamiento y control superior al 95% de su rango. Como ya se ha dicho, el FBA proporciona múltiples soluciones. Por lo tanto, se comprobó que los resultados obtenidos con la solución de suma de máximo flujo fueran representativos de todas las posibles soluciones que daban lugar a una optimización de la tasa de crecimiento.

1.10 El FBA predice una activación de las enzimas relacionadas con estrés oxidativo en células tratadas con MTF

Como se explica en el apartado anterior, empleando las combinaciones de flujos cuya suma fuese máxima y ponderando por su rango calculado mediante FVA, se realizó una comparativa entre el flujo de las células control y las células tratadas. De todos los candidatos evaluados, el FBA predice un flujo nulo en la catalasa en las células control con la excepción de las HCC1143, en las que el modelo predice una activación constitutiva de la catalasa. En las MDAMB231 y las

MCF7 tratadas con MTF, el modelo predice una activación de la catalasa mientras que en las CAMA1 no se observa ninguna alteración en el flujo de esta enzima (Figura 15).

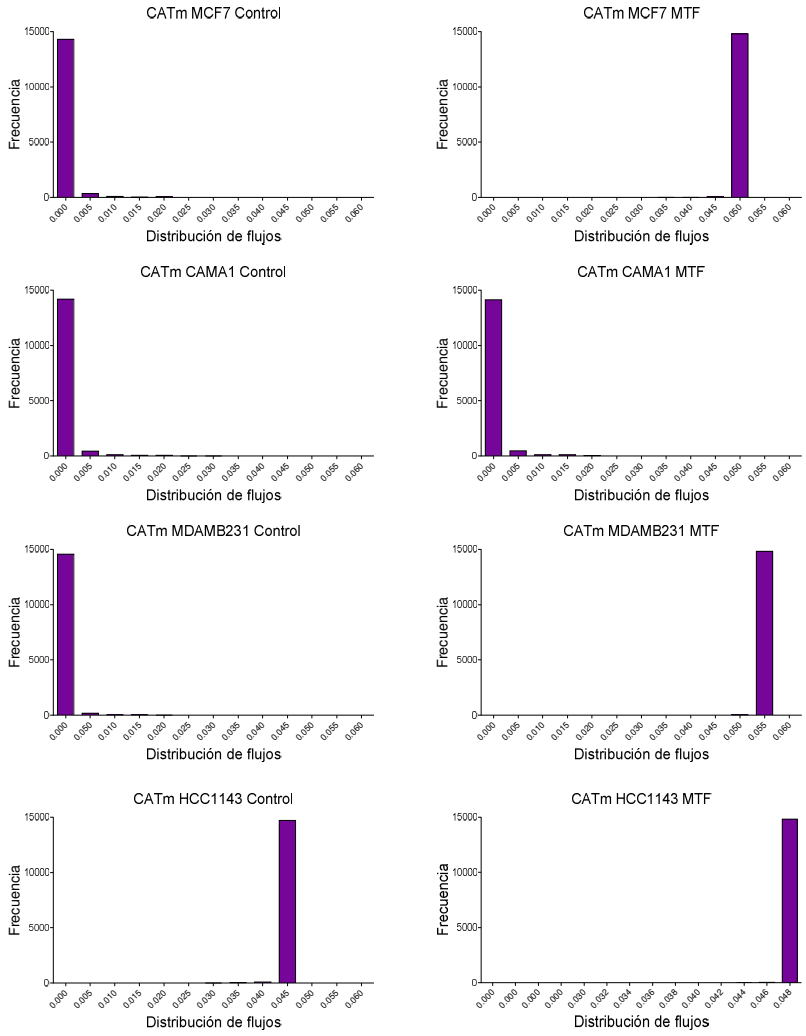


Figura 15: Distribución de posibles flujos de la catalasa (CATm) en células control y células tratadas con MTF. En el eje y se representa la frecuencia, entendida como el número de combinaciones posibles de flujo que proporciona el remuestreo por Monte Carlo, en el eje x el valor de flujo que toma la reacción.

Además, el modelo predice un incremento del flujo de la reacción SPODM en las MCF7 y las HCC1143 tratadas con MTF, pero no en las MDAMB231. Para las CAMA1, el modelo predice una alta actividad de la reacción SPODM tanto en el control como en las células tratadas con MTF (Figura 16).

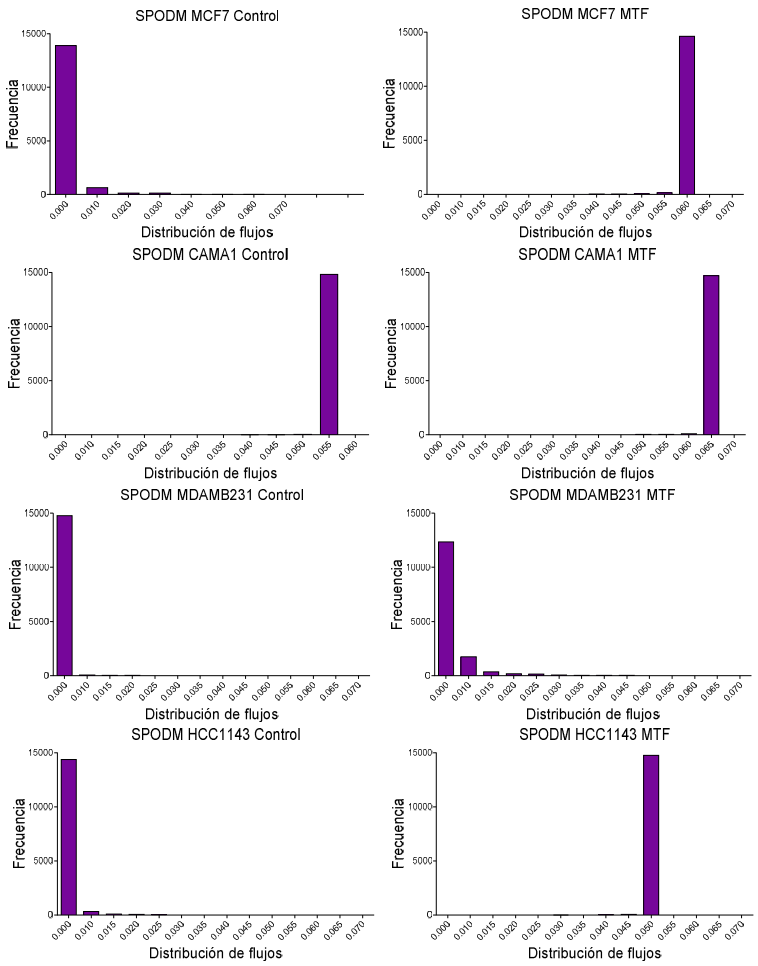


Figura 16: Distribución de posibles flujos de la SPODM en células control y células tratadas con MTF. En el eje y se representa la frecuencia, entendida como el número de combinaciones posibles de flujo que proporciona el remuestreo por Monte Carlo, en el eje x el valor de flujo que toma la reacción.

Por último, el modelo predice un incremento del flujo de la reacción de la óxido nítrico sintasa (NOS2) y, en consecuencia, un incremento en la producción de óxido nítrico (NO) en las MCF7 tratadas con MTF (Figura 17).

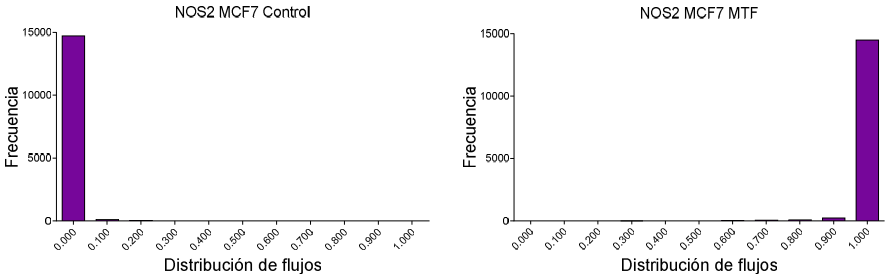


Figura 17: Distribución de posibles flujos de la NOS2 en MCF7 control y MCF7 tratadas con MTF. En el eje y se representa la frecuencia, entendida como el número de combinaciones posibles de flujo que proporciona el remuestreo por Monte Carlo.

Todas estas hipótesis se comprobaron en la distribución de posibles soluciones proporcionadas por el Monte Carlo, manteniéndose en todos los casos.

1.11 Las mediciones experimentales de la superóxido dismutasa confirman las predicciones hechas por el FBA

Para validar la hipótesis del modelo que sugiere una activación de la reacción SPODM por el tratamiento con MTF, se midió la actividad de la superóxido dismutasa tanto en las células control como en las células tratadas con MTF a 40 mM. Esta actividad se cuantificó mediante un ensayo enzimático basado en la medición de la absorbancia de la muestra a la que se le ha añadido un compuesto colorimétrico que reacciona con los aniones superóxido producidos por la SPODM. Con la excepción de las MCF7, las predicciones hechas por el FBA se confirman en los modelos experimentales. La actividad SPODM se calcula ponderando por el rango máximo y mínimo de absorbancia que puede tomar la muestra, es decir, por la absorbancia medida en una muestra que presente el 100% de actividad SPODM y una que no presente ninguna actividad SPODM. En las HCC1143, la actividad SPODM está ligeramente aumentada en las células tratadas con respecto a las células control. Por otro lado, las MDAMB231 son las que presentan la actividad SPODM más baja, como predecía el modelo, y las CAMA1 son las que presentan la actividad SPODM más alta tanto en el control como en las células tratadas, como también predice el modelo (Tabla 11).

Línea celular	Control Actividad SPODM (%)	Tratadas con MTF Actividad SPODM (%)
MCF7	96,44%	90,76%
CAMA1	99,01%	97,09%
MDAMB231	68,17%	49,82%
HCC1143	83,30%	86,44%

Tabla 11: Porcentaje de actividad SPODM medido experimentalmente en cada una de las seis líneas celulares de cáncer de mama tratadas con MTF y sin tratar.

1.12 Porcentaje de coincidencia entre el valor más frecuente del remuestreo y la primera solución que proporciona el FBA

El remuestreo por Monte Carlo, a pesar de proporcionar una visión global de las distribuciones de los flujos de las reacciones en las múltiples soluciones, presenta una limitación debido al alto tiempo de computación necesario para realizar el análisis, siendo imposible utilizarlo en series grandes. Para intentar solventar este problema, se estudió el porcentaje de coincidencia entre la combinación de flujos aportada por la función *optimizeCbModel* (un FBA estándar), considerablemente más eficaz en tiempo de computación, y los remuestreos por Monte Carlo en los datos de las líneas celulares tratadas y sin tratar. Para ello, se calculó el valor de coinci-

dencia C, siendo coincidentes (C=1) si cumplían el siguiente criterio: R> (M-0,01) & R< (M+0,01); siendo R el resultado de flujo del FBA y M el valor que toma la moda para ese flujo. El porcentaje de reacciones que presentan una coincidencia entre el valor más común en el remuestreo (que proporciona 14.480 posibles soluciones) y la solución que proporciona el FBA estándar es superior al 85% en todas las muestras (Tabla 12). Además, sólo existe discordancia en alguna muestra en 1.754 reacciones del total de 7.440 incluidas en la Recon2, lo que supone el 23% de las reacciones del modelo. Aquellas reacciones que presentan una discordancia en todas las muestras son en su mayoría reacciones de intercambio (33% del total de reacciones con discordancia en todas las muestras), que son reacciones con una categoría especial ya que no están limitadas por GPR al ser reacciones sumidero o reacciones de transporte.

Muestra	% de concordancia
MCF7 Control	86,84%
MCF7 MTF	86,25%
MCF7 RP	87,00%
CAMA1 Control	87,09%
CAMA1 MTF	86,73%
CAMA1 RP	86,73%
MDAMB231 Control	85,99%
MDAMB231 MTF	87,00%
MDAMB231 RP	86,04%
HCC1143 Control	87,06%
HCC1143 MTF	86,92%
HCC1143 RP	86,88%

Tabla 12: Porcentaje de concordancia para cada una de las muestras entre el valor más frecuente de flujo proveniente del remuestreo por Monte Carlo y el valor obtenido mediante el FBA estándar.

2. Aplicación del *Flux Balance Analysis* a datos de proteómica provenientes de muestras FFPE de tumores de cáncer de mama

2.1 Tasa de crecimiento tumoral predicha mediante FBA empleando datos de proteómica de muestras FFPE de tumores de mama

Los resultados obtenidos del FBA para los datos de proteómica fueron analizados con el fin de estudiar si el modelo reflejaba las diferencias conocidas en la tasa de crecimiento tumoral de los distintos subtipos de cáncer de mama. Se hallaron diferencias significativas en la tasa de crecimiento tumoral entre *ER-true* y *TN-like* (p-valor < 0,05). Como era de esperar, no existen diferencias significativas entre *TN-like* y TNBC (Figura 18).

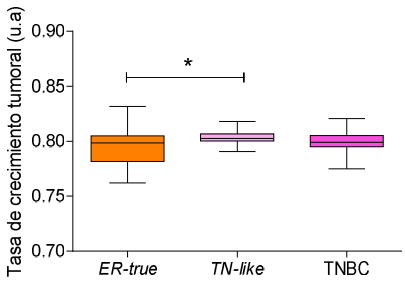


Figura 18: Tasa de crecimiento tumoral predicha por el FBA al introducir los datos de expresión de proteínas para *ER-true*, *TN-like* y TNBC (N= 51, 21 y 26 respectivamente, p-valor< 0,05). u.a.= unidades arbitrarias.

2.2 Predictor de recaída a distancia basado en las actividades de los flujos de las muestras tumorales

Una vez demostrado en los análisis en líneas celulares que la variación entre los resultados obtenidos mediante el remuestreo por Monte Carlo y la solución que aporta el FBA no es significativa (ver Resultados 1.12), se eligió para los siguientes análisis la solución que proporciona la función *optimizeCbmodel*, lo que permitía un gran ahorro en el tiempo de computación. De esta manera, se calcularon las actividades de los flujos para cada una de las rutas en los resultados obtenidos por un FBA estándar de los datos de proteómica de tumores de mama. Se encontraron diferencias significativas entre los subtipos descritos previamente en las rutas de interconversión de nucleótidos y detoxificación de radicales libres de oxígeno (ROS) (Figura 19).

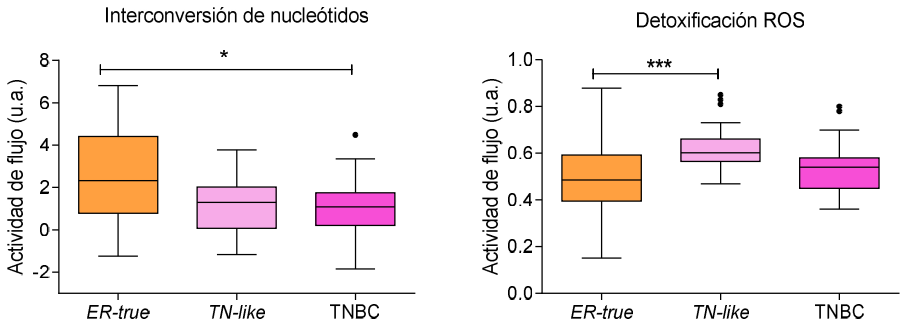


Figura 19: Actividades de flujo con diferencias significativas entre subtipos en los datos de proteómica de pacientes con cáncer de mama. u.a.= unidades arbitrarias

Además, se encontraron cinco actividades de flujo relacionadas con recaída a distancia (Tabla 13).

Actividad de flujo	p-valor paramétrico	FDR	Hazard Ratio
Metabolismo de tetrahidrobiopterina	0,0008	0,0472	2,817
Metabolismo de vitamina A	0,0027	0,0782	0,65
Metabolismo de beta-alanina	0,0299	0,444	3,23
Síntesis de coA	0,0396	0,444	0,504
Metabolismo de glioxilato y carboxilato	0,0496	0,444	2,07

Tabla 13: Actividades de flujo relacionadas con recaída a distancia en la cohorte de 96 pacientes de cáncer de mama. FDR: Tasa de falsos descubrimientos.

Se construyó un predictor de recaída a distancia basado en estas actividades de los flujos. Este predictor constaba de la actividad de las rutas del metabolismo de la vitamina A, la tetrahidrobiopterina y la beta-alanina y dividía a la población en un grupo de bajo riesgo y un grupo de alto riesgo (p-valor= 0,0032; HR= 6,52; 30%-70%) (Figura 20). El predictor se calcula mediante la fórmula $\sum_i w_i xi + 0,554$ (w=peso, x= valor de actividad de flujo, i= muestra), considerándose una muestra como de alto riesgo cuando el índice pronóstico es superior a -0,418 (Tabla 14).

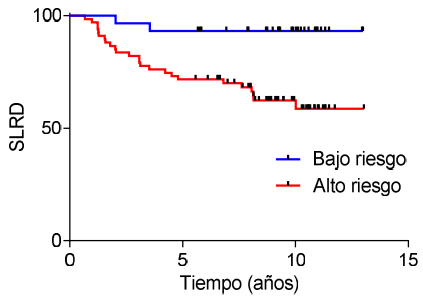


Figura 20: Predictor basado en la actividad de los flujos de las rutas de la vitamina A, la tetrahidrobiopterina y la beta-alanina en los datos de proteómica de pacientes de cáncer de mama. SLRD: Supervivencia libre de recaída a distancia.

Ruta metabólica	p-valor	Peso
Vitamina A	0,002	-0,468
Tetrahidrobiopterina	0,006	1,106
Beta-alanina	0,059	0,159

Tabla 14: Pesos asignados a cada una de las rutas metabólicas para el predictor de recaída.

Además, el valor pronóstico del predictor se mantiene en ambos subtipos de tumores ER+, siendo las diferencias estadísticamente significativas en el grupo *ER-true* (p-valor *ER-true* = 0,0179; p-valor *TN-like* = 0,064; p-valor TNBC = 0,364) (Figura 21).

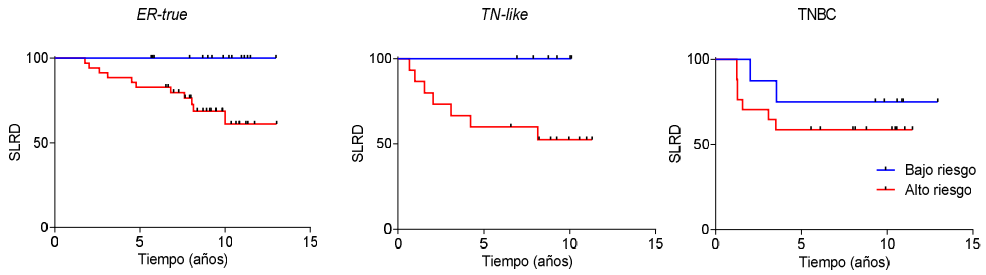


Figura 21: Predictor basado en las actividades de los flujos de la ruta de la vitamina A, la beta alanina y la tetrahidrobiopterina por subtipo molecular. SLRD: Supervivencia libre de recaída a distancia.

El modelo de regresión multivariante de Cox muestra que este predictor proporciona información adicional con respecto a los datos clínicos (Tabla 15).

Análisis multivariante	p-valor
T	0,338
N	0,015
G	0,061
Predictor	0,004

Tabla 15: Regresión de Cox comparando el predictor basado en las actividades de los flujos de la beta-alanina, tetrahidrobiopterina y vitamina A. T= tamaño tumoral, N= afectación ganglionar, G= grado histológico.

Se han descrito diferencias a nivel de metabolismo en los tumores TNBC respecto a los ER+. Por este motivo, se construyó un predictor exclusivamente para este tipo de tumores. Esta nueva firma se compone de las actividades de los flujos de las rutas de la glucolisis y el metabolismo del glutamato y divide a la población de TNBC en un grupo de alto y un grupo de bajo riesgo (p-valor= 0,106; HR= 4,60; 30-70%), a pesar de no ser las diferencias estadísticamente significativas (Figura 22). El predictor se calcula mediante la fórmula $\sum_i w_i\, x_i - 2{,}697$, (w= peso, x= valor de actividad de flujo, i= muestra), clasificando a la muestra en el grupo de alto riesgo si el índice pronóstico es mayor de -0,690 (Tabla 16).

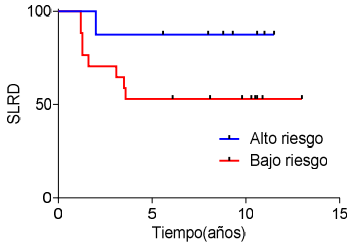


Figura 22: Predictor en tumores TNBC basado en las actividades de los flujos de las rutas de glucolisis y metabolismo del glutamato. SLRD: Supervivencia libre de recaída a distancia.

Ruta metabólica	p-valor	Peso
Glucolisis	0,031	0,892
Metabolismo de glutamato	0,033	-0,491

Tabla 16: Pesos asignados a cada actividad de flujo contenida en el predictor para los tumores TNBC.

En este caso, el análisis multivariante no presenta significación (Tabla 17).

Análisis multivariante	p-valor
T	0,977
N	0,733
G	0,514
Predictor	0,206

Tabla 17: Análisis multivariante de Cox comparando el predictor basado en las actividades de los flujos de glucolisis y metabolismo del glutamato en TNBC. T= tamaño tumoral, N= afectación ganglionar, G= grado histológico.

3. Estudio de asociación de datos de metabolómica con los resultados obtenidos en el *Flux Balance Analysis* y con datos de expresión génica en una cohorte de pacientes de cáncer de mama

Con el fin de estudiar la relación entre los datos de expresión génica, de metabolómica y las actividades de los flujos del FBA, se analizaron los datos provenientes del trabajo de Terunuma *et al.*, en el que estaban disponibles datos de metabolómica y de *arrays* de expresión génica para la misma cohorte de pacientes de cáncer de mama (112). Esta cohorte está formada por 67 pacientes, 34 de ellos ER-, de los cuales 14 están especificados como TNBC, y 33 ER+. Las muestras provenían de tejido fresco y se habían cuantificado los metabolitos mediante espectrometría de masas y medido la expresión génica empleando *microarrays* GeneChip Human Gene 1.0 ST de Affymetrix.

3.1 Análisis basados en los datos de metabolómica

Se construyó un predictor de supervivencia global usando los datos de metabolómica. Este predictor se compone de los metabolitos: glutamina, 2-hidroxipalmitato, deoxicarnitina, butirilcarnitina y glicerofosforilcolina (p-valor= 0,003; HR= 0,34; 50-50%) (Figura 23). La fórmula para calcular el predictor es $\sum_i w_i + x_i - 15{,}500$ y una muestra se considera de alto riesgo cuando el índice pronóstico es mayor de 0,246 (Tabla 18).

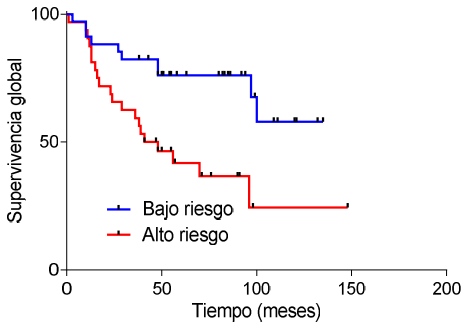


Figura 23: Predictor basado en datos de metabolómica.

Metabolito	p-valor	Peso
2-hidroxipalmitato	0,0009	0,142
Deoxicarnitina	0,0030	0,250
Glutamina	0,0035	-0,0760
Butirilcarnitina	0,0041	0,261
Glicerofosforilcolina	0,0364	0,240

Tabla 18: Pesos asignados a cada uno de los metabolitos en el predictor.

El análisis multivariante confirma que el predictor proporciona información adicional a los parámetros clínicos (Tabla 19).

Análisis multivariante	p-valor
T	0,863
N	0,014
G	0,246
Predictor metabolitos	0,018

Tabla 19: Modelo de regresión multivariante de Cox comparando el predictor de supervivencia global basado en los datos de metabolitos. T= tamaño tumoral, N= estatus ganglionar, G= grado histológico.

Se construyó una red formada únicamente por los datos de cuantificación de los metabolitos usando los MGP. La base de datos de metabolómica contaba con datos de la medición de 536 metabolitos diferentes que tras los criterios de calidad aplicados se reducen a 237 metabolitos. Posteriormente, se asignó una ruta metabólica mayoritaria a cada una de las ramas de la red mediante análisis ontológico con IMPaLA. Al igual que se venía observando en las redes provenientes de expresión génica, la red agrupaba a los metabolitos por rutas metabólicas (Figura 24).

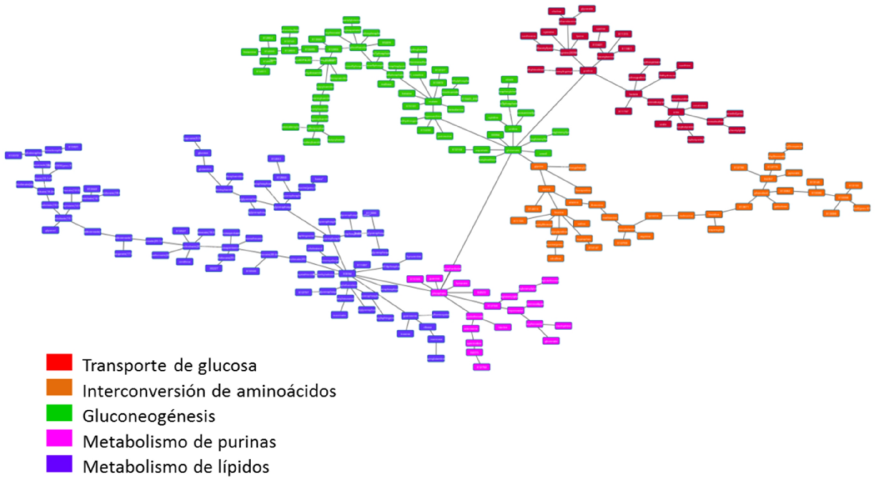


Figura 24: Red de metabolitos proveniente de los datos publicados por Terunuma *et al.*

Se calculó la actividad de los nodos de la misma manera en la que se calculaban en las redes de expresión génica y se realizaron comparaciones entre los tumores ER+ y ER-. Existían diferencias significativas en el metabolismo de los lípidos y en el metabolismo de las purinas entre estos dos grupos (p< 0,05) (Figura 25).

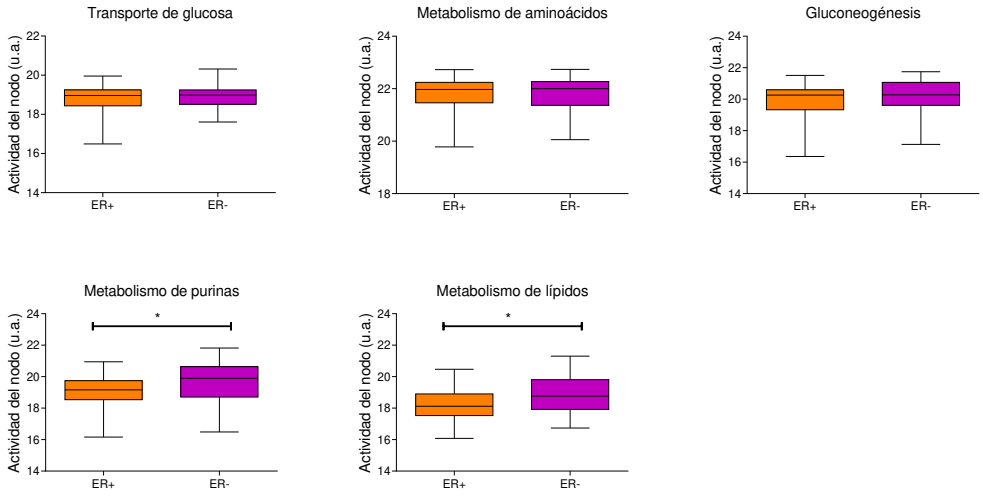


Figura 25: Actividad de los nodos de la red compuesta por metabolitos. u.a.= unidades arbitrarias.

Además, la actividad del nodo de metabolismo de lípidos presentaba valor pronóstico en esta serie (p= 0,0452; HR= 0,47; 50-50%) (Figura 26, Tabla 20). El predictor puede calcularse mediante la fórmula $\sum_i w_i + x_i - 8{,}689$, considerándose de alto riesgo una muestra con un índice pronóstico mayor de -0,0629.

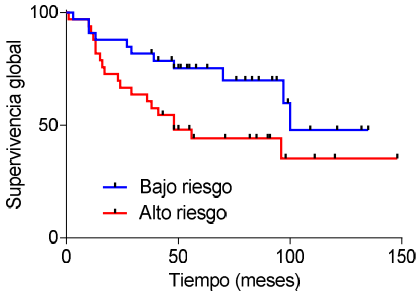


Figura 26: Predictor basado en la actividad del nodo del metabolismo de lípidos.

Nodo	p-valor	Peso
Metabolismo de lípidos	0,008	0,468

Tabla 20: Pesos asignados al predictor compuesto por la actividad del nodo de metabolismo de lípidos.

Sin embargo, el análisis multivariante no muestra que el predictor aporte información significativa frente a los datos clínicos (Tabla 21).

Análisis multivariante	p-valor
T	0,732
N	0,030
G	0,464
Predictor actividad nodo metabolismo lípidos	0,141

Tabla 21: Análisis multivariante de Cox comparando el predictor basado en la actividad del nodo de metabolismo de lípidos con los datos clínicos. T= tamaño tumoral, N= estatus ganglionar, G= grado histológico.

3.2 Combinación de datos de expresión génica con datos de metabolómica

Por otro lado, se combinaron los datos de expresión génica con los datos de metabolómica en una única red. Debido a las diferencias entre los dos tipos de datos, la mayoría de los metabolitos se agrupaban aislados en un único nodo de la red. Sin embargo, había unos pocos metabolitos que se mezclaban con los datos de expresión génica (Figura 27 A). Esta red se caracterizó en base a la función mayoritaria de los genes de cada rama de la red. La red estaba compuesta de once nodos funcionales y un duodécimo nodo que agrupaba a la mayoría de los metabolitos (Figura 27 B).

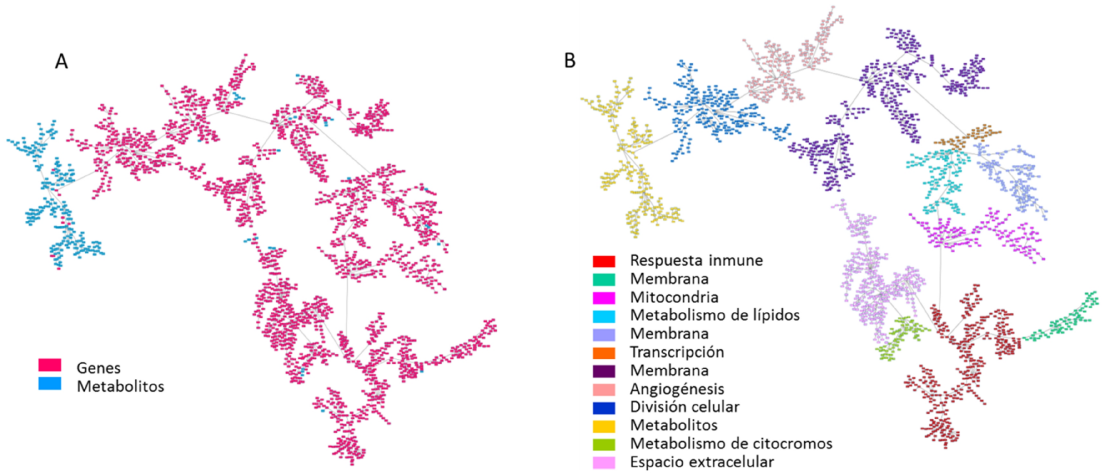


Figura 27: A. Red combinada de datos de expresión génica y datos de metabolómica. B. Red combinada de datos de expresión génica y datos de metabolómica caracterizada funcionalmente.

Una vez establecidas las funciones mayoritarias, se estudió si existía alguna asociación descrita entre los metabolitos que se incluían en los nodos compuestos por genes y la función asignada a su correspondiente nodo. Existían relaciones previas descritas en la bibliografía en el caso de cuatro de los 20 metabolitos identificados embebidos en los nodos de genes: succinato, citidina, histamina y 1,2-propanediol. Todas las asociaciones están recogidas en la tabla 22.

Metabolito	Función del nodo al que pertenece	Relación descrita	Referencia
Succinato	Respuesta inmune	Aumento de la respuesta inmune, inducción de la producción de IL-1b, promueve la respuesta inmune adaptativa.	(124, 125)
Citidina	Respuesta inmune	5-aza-2'-deoxi-citidina potencia la respuesta inmune antitumoral, papel en respuesta inmune innata.	(126)
Histamina	Angiogénesis	Promueve la angiogénesis a través de la producción de VEGF.	(127)
1,2-propanediol	Angiogénesis	Modula al sistema inmune a través de S1P que promueve la angiogénesis y la proliferación. 14C-sulfoquinovosilacilpropanediol es un fármaco antiangiogénico.	(128)

Tabla 22: Descripción de la asociación de los metabolitos con la función de sus correspondientes nodos.

3.3 Combinación de datos de metabolómica con datos de actividades de los flujos

Se calculó el FBA para estos datos como se ha explicado en secciones previas. Brevemente, se utilizaron los datos de expresión génica para las 67 muestras tumorales, se resolvieron las GPR y se introdujeron los datos de expresión en el modelo mediante el *E-flux* modificado. Por último, se calculó el FBA teniendo como función objetivo la reacción de biomasa como representante de la tasa de crecimiento tumoral. No existían diferencias significativas en la tasa de crecimiento tumoral entre ER+ y ER- (Figura 28).

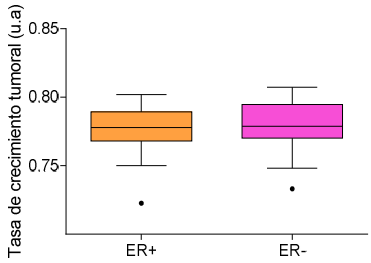


Figura 28: Tasa de crecimiento tumoral predicha mediante FBA para los pacientes de esta cohorte. No existen diferencias significativas entre ER+ y ER-. u.a.= unidades arbitrarias.

A nivel de actividades de los flujos se encontraron diferencias significativas entre ER+ y ER- en metabolismo de glicerofosfolípidos, metabolismo de fosfatidilinositol, ciclo de la urea, metabolismo de propanoato, catabolismo de pirimidinas y detoxificación de ROS (Figura 29).

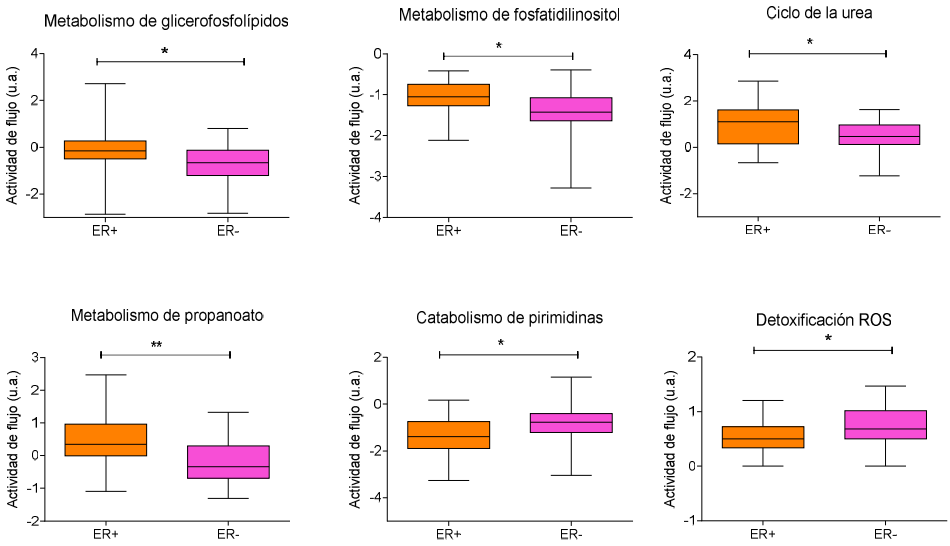


Figura 29: Actividades de los flujos diferenciales entre ER+ y ER-. u.a.= unidades arbitrarias.

Además, el metabolismo de la glutamina y el metabolismo de la alanina y el aspartato tienen valor pronóstico en esta cohorte (p-valor= 0,0243; HR= 0,411; 50-50%) (Figura 30, Tabla 23). La fórmula para calcular el predictor es $\sum_i w_i + x_i + 0{,}681$ y se clasifica una muestra en el grupo de alto riesgo cuando su índice pronóstico es mayor de -0,0589.

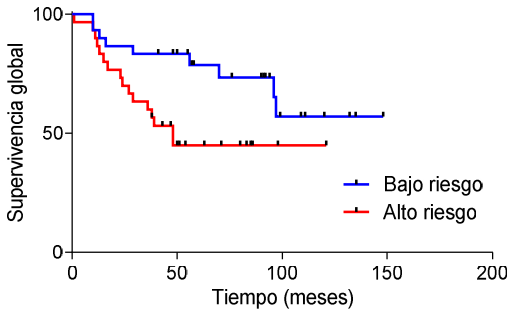


Figura 30: Predictor basado en las actividades de los flujos del metabolismo de la glutamina y de la alanina y el aspartato.

Actividad de flujo	p-valor	Peso
Metabolismo del glutamato	0,005	-0,569
Metabolismo de alanina y aspartato	0,040	0,314

Tabla 23: Pesos asignados según el predictor a cada una de las actividades de los flujos.

En este caso, el análisis multivariante sí muestra que el predictor aporta información sobre los parámetros clínicos (Tabla 24).

Análisis multivariante	p-valor
T	0,489
N	0,058
G	0,351
Predictor actividades de flujo	0,028

Tabla 24: Modelo de regresión multivariante de Cox comparando el predictor basado en las actividades de los flujos. T= tamaño tumoral, N= afectación ganglionar, G= grado histológico.

Se calcularon las actividades de los flujos para cada una de las rutas metabólicas definidas en la Recon2 como se ha descrito anteriormente y se combinaron mediante MGP con los datos de metabolómica para estudiar su asociación. La red mezclaba los dos tipos de datos, aunque situaba a las actividades de los flujos en la periferia de la red (Figura 31 A).

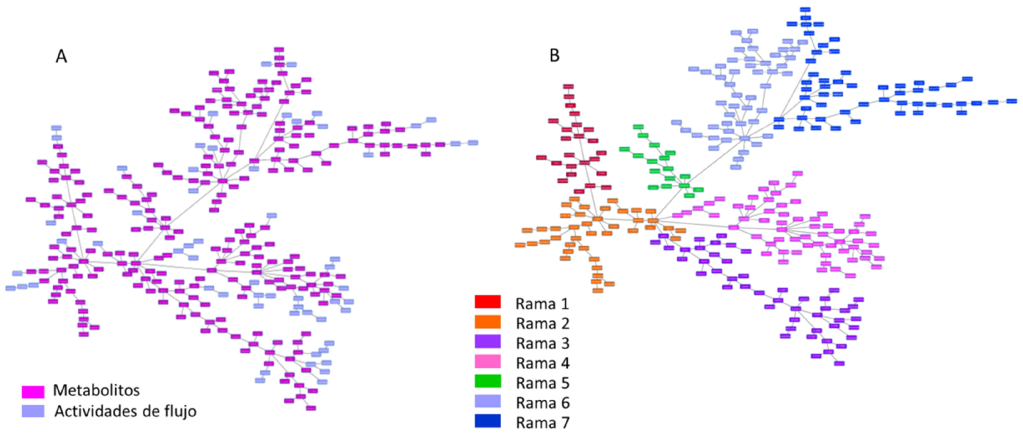


Figura 31: A. Red resultante de combinar los datos de metabolómica con las actividades de los flujos calculadas para cada una de las rutas metabólicas definidas en la Recon2. B. Red de metabolitos y actividades de los flujos dividida por ramas.

Esta red se dividió en ramas y se estudió la asociación de los metabolitos pertenecientes a cada rama con las actividades de los flujos correspondientes a esa rama de la red (Figura 31 B). Se vio que existía una coherencia entre ambos tipos de datos, conteniendo cada rama metabolitos pertenecientes a la ruta indicada en la actividad de flujo. Por ejemplo, en la rama 1 se encuentra la actividad de flujo correspondiente a la glucolisis y en esta misma rama aparecen tres metabolitos que intervienen en esa ruta según la información de la base de datos IMPaLA. En el caso del metabolismo de vitaminas no fue posible cuantificar el número de metabolitos relacionados debido a que la etiqueta ontológica en IMPaLA es “metabolismo de vitaminas y co-factores” mientras que en la Recon2 distinguen entre cada una de las vitaminas, etiquetando las actividades de los flujos con etiquetas más concretas como por ejemplo “metabolismo de vitamina B6” (Anexo 1).

4. Interfaz FLUX para facilitar la realización del *Flux Balance Analysis*

Para realizar el FBA con la librería COBRA Toolbox son necesarios conocimientos básicos de programación en MATLAB. Con el fin de hacer accesible este análisis, se creó una interfaz amigable basada en las funciones y el código de COBRA Toolbox mediante la GUIDE de MATLAB en la que no es necesario utilizar lenguaje de programación. Esta interfaz se inicia abriendo MATLAB y escribiendo “FLUX” en la consola. Una vez iniciada, la interfaz permite llevar a cabo todos los pasos necesarios para realizar un FBA. Se puede importar un modelo escogido, importar un archivo que contenga las reglas GPR para ese modelo, fijar la función objetivo en la reacción de biomasa propuesta en la Recon2 y, finalmente, calcular el FBA. Además, esta aplicación permite calcular el FVA y realizar un análisis de reacciones esenciales mediante *knockouts*. Para comprobar que se ha realizado con éxito cada uno de los procesos, la interfaz está

diseñada para que aparezca un mensaje de confirmación en un recuadro blanco en cada uno de los pasos. En el caso del FBA, el FVA y el análisis de reacciones esenciales por KO, los resultados se exportan a un archivo en formato *.txt* a la carpeta de trabajo con el nombre de "Results", "minFluxResults" y "maxFluxResults", y "KOs" respectivamente (Figura 32). Una vez terminado el análisis la interfaz dispone de un botón con un mensaje de salida. El código de esta interfaz se suministra en el Anexo 2.

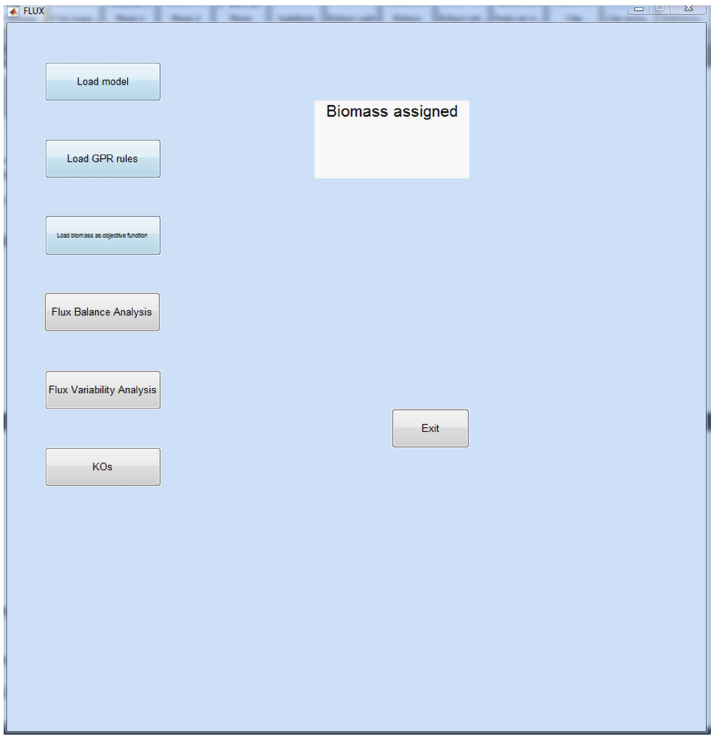


Figura 32: Aplicación FLUX creada mediante la GUIDE de MATLAB para realizar FBA, FVA y análisis de *knockouts* sin necesidad de utilizar lenguaje de programación.

DISCUSIÓN

DISCUSIÓN

La reprogramación del metabolismo es uno de los procesos principales en cáncer. En concreto, en cáncer de mama se han descrito diferencias en el metabolismo entre los subtipos moleculares (30, 41). Por tanto, los fármacos con dianas metabólicas podrían ser un complemento útil a la quimioterapia clásica en el tratamiento de pacientes sin terapia dirigida establecida, como es el caso de los TNBC, o en aquellos que desarrollen resistencias. El objetivo principal de este trabajo es la caracterización de las alteraciones metabólicas en tumores y líneas celulares de cáncer de mama y el estudio de la respuesta de estas líneas celulares a fármacos que afectan a procesos metabólicos. Para ello se han empleado modelos computacionales como MGP y FBA.

1. Experimentos de perturbación

En este trabajo se emplearon datos de proteómica y modelos computacionales, con el fin de caracterizar la respuesta que se produce en distintas líneas celulares de cáncer de mama al tratamiento con fármacos que afectan al metabolismo, como son la MTF y la RP, y se demuestra la utilidad de este tipo de aproximaciones computacionales para estudiar los efectos y mecanismos de acción de estos fármacos. Hasta dónde nosotros sabemos, este es el primer estudio que combina proteómica y análisis computacionales para estudiar los mecanismos de acción de fármacos, aunque existe un estudio previo en el que se empleó FBA para proponer nuevas dianas terapéuticas en células de cáncer de ovario y estudiar sinergias de varios fármacos (129).

1.1 Respuesta de las líneas celulares de cáncer de mama a fármacos que afectan al metabolismo

En trabajos previos nuestro grupo demostró que existían diferencias significativas en el metabolismo de la glucosa entre tumores de los distintos subtipos de cáncer de mama. Además se vio que las líneas celulares TNBC mostraban una producción de lactato más alta y un consumo de glucosa mayor con respecto a las ER+ (41). Estas alteraciones metabólicas sugieren la posibilidad de usar fármacos contra dianas metabólicas en pacientes de cáncer de mama y lleva a pensar que la respuesta será heterogénea entre los diferentes subtipos.

Efectivamente, las curvas dosis-respuesta confirman que la respuesta de las células de cáncer de mama a estos fármacos es heterogénea. Con la administración de MTF se observa un marcado efecto en la viabilidad celular, siendo las CAMA1 las más resistentes a MTF y las MDAMB468 las que presentan una mayor sensibilidad. Sin embargo, no se pudieron asociar estas diferencias al subtipo. En el caso de la RP, la respuesta sí es subtipo-dependiente, siendo

más efectiva en las líneas celulares ER+ que en las TNBC. Estos resultados tienen su reflejo en la clínica, ya que un derivado de la RP, el *everolimus*, se utiliza en el tratamiento de cáncer de mama ER+ en estadios avanzados (50).

1.2 Genotipado de polimorfismos

Con el fin de explicar esta heterogeneidad en la respuesta, se empleó un *array* de SNP para estudiar los polimorfismos que pudiesen estar involucrados. La caracterización de los SNP sugería que la elevada sensibilidad a MTF observada en las MDAMB468 podría deberse a la presencia del polimorfismo rs2282143 en el transportador *SLC22A1*, que está relacionado con una disminución del aclaramiento de la MTF (https://www.pharmgkb.org/). Por otro lado, el polimorfismo rs628031 en *SLC22A1*, previamente asociado con una pobre respuesta a MTF, está presente en homocigosis en las MCF7 y las HCC1143 (https://www.pharmgkb.org/).

Respecto a la RP, todas las líneas celulares ER+ presentan en heterocigosis el polimorfismo rs1045642 en el gen *ABCB1*, que no ha sido relacionado previamente con la respuesta a RP. En *CYP3A4*, rs2740574, relacionado con altos requerimientos de *sirolimus*, está presente en heterocigosis en las MDAMB468 (https://www.pharmgkb.org/).

Estos resultados sugieren un papel del polimorfismo rs2282143 en *SLC22A1* en la sensibilidad a MTF en las MDAMB468. En el caso de los demás polimorfismos debería estudiarse su efecto en profundidad con otros experimentos, como por ejemplo ensayos funcionales.

1.3 Proteómica en líneas celulares de cáncer de mama tratadas y sin tratar

Se realizaron experimentos de perturbación en las seis líneas celulares de cáncer de mama empleando una dosis sub-letal de MTF y RP, seguidos de experimentos de proteómica. Es relevante el uso de datos de proteómica en lugar de expresión génica, como es habitual, ya que el trabajo de Sacco *et al*., en el que comparan transcriptoma y proteoma de células MCF7 tratadas con MTF y sin tratar, encuentran un menor reflejo de los cambios producidos por el tratamiento a nivel de ARNm, siendo la correlación media entre los datos de expresión génica y proteínas de tan sólo el 0,45 (130).

Con los nuevos avances en la tecnología y análisis de la proteómica, se consiguieron detectar y cuantificar del orden de 7.000 proteínas por muestra, lo que cubre casi el total del proteoma de la célula. Por tanto, desaparece la desventaja con respecto a la genómica en cuanto a la cobertura del número de proteínas analizables, poniéndose la proteómica a la altura del análisis de la expresión génica.

1.4 Caracterización de las diferencias a nivel proteico debidas al tratamiento

Se caracterizaron las diferencias en la expresión de proteínas entre las células control y las células tratadas con MTF y RP. Algunas de estas proteínas diferenciales coincidían con las interacciones descritas en la *Comparative Toxicogenomics Database*, como por ejemplo un aumento en la expresión de SIRT2 (sirtuina 2) y HTATIP2 (proteína interactiva de Tat 2, asociada con cáncer de cérvix), y una disminución de la expresión de SIRT5 (sirtuina 5), PPP4R2 (subunidad regulatoria 2 de la proteinfosfatasa 4) y MYD88 (adaptador de transducción de señal inmune innata) debidas a la administración de MTF. Ya había sido previamente descrito un aumento en la expresión de *SIRT2* inducida por MTF (131). *SIRT2* tiene un importante papel en el proceso de inflamación, promueve la gluconeogénesis y aumenta la defensa ante ROS (132). Esto último concuerda con las predicciones hechas por nuestro modelo. Además, la administración de MTF conlleva una disminución de la expresión de *SIRT5* (131). Está disminución está también relacionada con las diferencias de flujo observadas en el modelo, ya que está descrito que SIRT5 está implicada en la regulación de la SPODM (133), coincidiendo con las predicciones del FBA acerca de la activación de la SPODM en respuesta a estrés oxidativo. Por otro lado, el MGP establece que existe una disminución de la actividad del nodo de procesamiento de ARNm, replicación del ADN, mitocondria B y unión a ATP.

También se han caracterizado diferencias asociadas al tratamiento con RP, como por ejemplo un incremento en la expresión de *NR3C1* (receptor nuclear de la subfamilia 3 miembro 1 de grupo C) y *RPS27L* (proteína ribosomal *S27-like*) y una disminución en *CKS1B* (subunidad 1B regulatoria de la proteína quinasa CDC28), *COL1A1* (cadena alfa 1 de colágeno tipo I), *IGFBP5* (proteína de unión al factor de crecimiento similar a la insulina 5), *SCD* (desaturasa esteraoil-CoA), mTOR y *CDK4* (quinasa dependiente de ciclina 4), previamente descritas (134). La inhibición de CDK4/6 suprime la progresión del ciclo celular en modelos celulares ER+/HER2- y complementa la actividad de los estrógenos (135). El tratamiento con RP también conlleva la disminución de la expresión del ARNm de *CSK1B* (136). Por otro lado, la bajada en la expresión de *CSK1B* promueve la apoptosis en células de cáncer de mama (137). Además, la RP produce una disminución de la expresión del ARNm de *KIFC1* (138), cuya sobreexpresión está descrito que promueve la proliferación (139). También está descrito que el tratamiento con RP produce un incremento en la actividad de la proteína NR3C1 (140). *NR3C1* codifica al receptor de glucocorticoides, que está implicado en respuesta inflamatoria y que tiene un efecto anti-proliferativo (141). Además, la RP promueve la unión de TP73 al promotor de RPS27L, diana directa de p53, y como consecuencia promotor de apoptosis (142). Por último, la RP inhibe la expresión del ARNm de *SCD* a través de TP73 (143). El 17-β-estradiol induce la expresión de *SCD* y la regula-

ción de la composición de lípidos celulares en células ER+ y es necesario para la proliferación inducida por estrógenos (144). En conjunto, todos estos resultados sugieren un efecto antiproliferativo de la RP. Por último, la RP induce la disminución de los niveles de mTOR (145-147). El MGP sugiere que la RP provoca una disminución en la actividad del nodo de procesamiento de ARNm. Además, la actividad de los nodos metabolismo A y metabolismo B predicen la respuesta a RP. Esto podría ser debido a que en las líneas celulares resistentes al fármaco el metabolismo sigue siendo un proceso activo.

1.5 Estudio del ciclo celular mediante citometría de flujo

El uso de la proteómica combinada con los análisis de ontología permitió explorar los cambios en la expresión de proteínas entre células control y células tratadas, que sugerían que estos fármacos afectaban al ciclo celular. Parecía lógico pues estudiar el ciclo celular a través de citometría de flujo para confirmar esta hipótesis. Se confirmó un arresto del ciclo celular en fase G2/M para las células estudiadas tratadas con MTF, con la excepción de las CAMA1 en donde la MTF presenta un efecto muy reducido sobre la viabilidad celular. En el caso de las ER+ tratadas con RP (pero no en la línea TNBC estudiada) se observó un arresto del ciclo celular en G0/G1, coincidente con los datos a nivel de proliferación. Está establecido que mTOR (diana de la RP) controla la progresión del ciclo celular mediante S6K1 y 4E-BP1 (148). Además, ya ha sido previamente descrito un arresto en G0/G1 en células MCF7 tratadas con RP (51). Por lo tanto, tanto la MTF como la RP poseen efectos citostáticos en líneas celulares de cáncer de mama, que conllevan una reducción en su viabilidad, combinado con una disrupción del ciclo celular. Sin embargo, esta respuesta es diversa entre las distintas líneas celulares.

1.6 Predicciones del crecimiento tumoral mediante el FBA y estudio de las actividades de los flujos en las líneas celulares

El FBA es un método computacional que tradicionalmente se usa en biotecnología para optimizar el crecimiento de microorganismos. Recientemente, con la aparición de reconstrucciones más completas del metabolismo humano, ha empezado a utilizarse en otros ámbitos, como el estudio de los glóbulos rojos (84) o el estudio del efecto Warburg (85).

En este trabajo se ha desarrollado un modelo metabólico que usa datos de expresión de proteínas para predecir crecimiento tumoral. En trabajos anteriores se describieron modelos metabólicos en cáncer que utilizaban datos de expresión génica y, en algunos casos, modelos reducidos del metabolismo (19, 43, 44). Nuestro modelo, sin embargo, usa la reconstrucción completa del metabolismo Recon2 y datos de proteómica con el fin de mejorar la exactitud de las predicciones. Hemos validado el modelo mediante la comparación con datos de crecimien-

to experimentales en células ER+ (MCF7 y T47D) y TNBC (MDAMB231 y MDAMB468). Esta aproximación permite la formulación de nuevas hipótesis y proporciona una visión global del metabolismo.

Las predicciones del modelo son coherentes con los cambios detectados en la viabilidad celular en las células tratadas con MTF. Hemos explorado el flujo global de cada ruta calculando las actividades de los flujos para identificar aquellas rutas metabólicas que presentan un comportamiento diferencial entre las células tratadas con MTF y las células control. Las rutas relacionadas con la respuesta a MTF son el metabolismo del glutamato y el metabolismo del piruvato. Las rutas metabólicas relacionadas con la respuesta a RP son el metabolismo de valina, leucina e isoleucina y el metabolismo del colesterol. A pesar de que es difícil realizar comparaciones entre patrones de flujo, el nuevo método propuesto basado en las actividades de los flujos podría ser una aproximación útil para solventar este problema y caracterizar diferencias en los patrones de flujo entre diferentes condiciones.

1.7 Remuestreo por Monte Carlo

Una de las principales limitaciones del FBA es la existencia de múltiples combinaciones de flujos posibles para un mismo valor óptimo de la función objetivo, lo que matemáticamente se conoce como un sistema de ecuaciones compatible indeterminado (87). Algunas de las aproximaciones propuestas para abordar este problema son un remuestreo de posibles soluciones mediante Monte Carlo (106) o una aproximación llamada *geometric FBA* (149). El *geometric FBA* consiste en varias iteraciones; la primera de ellas está basada en eliminar ciclos internos de reacciones debido a que son ineficaces y encontrar el centro del poliedro de posibles soluciones de manera aproximada; durante la segunda iteración se busca la combinación de flujos a menor distancia del centro que proporcione un valor óptimo para la función objetivo; durante la última iteración se acota esta área hasta llegar a una combinación de flujos única.

Sin embargo, estos métodos proporcionan una solución basada en métodos matemáticos que no tiene por qué ser la más representativa a nivel biológico. Existe otro método consistente en minimizar la suma de los flujos ponderado por un vector de pesos calculado a partir de los datos de expresión génica llamado *E-Fmin*. Este método se basa en la premisa de que la célula optimizará sus recursos sintetizando la menor cantidad de enzimas posibles (99).

En este trabajo hemos medido la cantidad de proteína experimentalmente y, mediante las GPR, se calcula la abundancia de cada enzima. Asumiendo que la célula utiliza todas las enzimas que sintetiza, nuestra propuesta es seleccionar la solución proporcionada por el remues-

treo por Monte Carlo cuya suma de flujos sea mayor, que es la solución en la que la célula emplea la máxima cantidad de proteínas medidas experimentalmente.

1.8 Predicción de la activación de enzimas relacionadas con respuesta a estrés oxidativo en células tratadas con MTF y validación experimental mediante la medición de la actividad de la superóxido dismutasa

Mediante el FVA y el remuestreo por Monte Carlo, se predijo una activación de las enzimas relacionadas con la respuesta a estrés oxidativo en las células tratadas con MTF. Se ha descrito una activación de la catalasa y de la SPODM por MTF en otros escenarios (150, 151) y, como se ha mencionado previamente, en la mayoría de los casos concuerda con diferencias observadas en la expresión de proteínas, aunque la relación no es siempre directa. Esto es debido a que los flujos de una reacción no dependen exclusivamente de la expresión de la proteína, sino que también están condicionados por los flujos de las reacciones de alrededor. Además, los flujos de la catalasa y de la SPODM parecen estar relacionados con la viabilidad. Por ejemplo, en las CAMA1 tratadas con MTF no se observa un incremento en el flujo de la catalasa, quizás debido al reducido efecto que tiene la MTF en la viabilidad de estas células. Algunas de estas predicciones han sido verificadas experimentalmente en el ensayo de actividad de la SPODM. En general, las mediciones de la actividad de la SPODM son consistentes con las predicciones hechas por el FBA. Las variaciones entre las predicciones del FBA y las mediciones experimentales de la SPODM pueden deberse al hecho de que el FBA sólo tiene en cuenta la información sobre las rutas metabólicas, obviando el resto de procesos celulares.

1.9 Predicción de la activación de la óxido nítrico sintasa en MCF7 tratadas con MTF

Por otro lado, el modelo predice un aumento en el flujo de la NOS2 en las MCF7 tratadas con MTF, como había sido previamente descrito en ratas diabéticas (152). Un aumento en el flujo de la NOS2 implica una mayor producción de NO. Un incremento en la concentración de NO se asocia con la activación de procesos apoptóticos y efectos citostáticos en células tumorales, mientras que bajas concentraciones de NO se asocian con supervivencia celular y proliferación (153). Este aumento de la NOS2 podría estar relacionado con la reducción en la proliferación observada en las MCF7 tratadas con MTF. El hecho de que este aumento de NO se haya predicho sólo en las MCF7 podría deberse a la heterogeneidad en los mecanismos de respuesta contra este tipo de fármacos y podría estar asociado con las diferencias vistas en la viabilidad celular entre las diferentes líneas. Es reseñable que, a pesar de no haber proporcionado al modelo ninguna información acerca de la abundancia de la NOS2, el modelo es capaz de refle-

jar diferencias a nivel de flujo en este proceso, sugiriendo que ambas aproximaciones (proteómica y FBA) proporcionan información complementaria.

1.10 Resumen de los resultados establecidos para líneas celulares tratadas con MTF

Las actividades de los nodos calculadas a partir del MGP para los nodos de mitocondria y unión a ATP sugieren que la MTF ejerce su acción en la mitocondria, hecho ampliamente establecido (46). Como se deduce de los datos del FBA, parece provocar un incremento de las enzimas asociadas a respuesta a estrés oxidativo. Además, en las MCF7 se predice un incremento en la NOS2. Por último, la elevada susceptibilidad de las MDAMB468 a la MTF podría deberse a la presencia del polimorfismo en el transportador *SLC22A1.* Como consecuencia de estos eventos, la MTF causa un efecto heterogéneo en la proliferación celular, consistente con un arresto en G2/M.

1.11 Resumen de los resultados establecidos para líneas celulares tratadas con RP

La RP ejerce un marcado efecto en la proliferación celular de las células ER+, mediante un arresto en G0/G1, como ha sido previamente descrito (154). Esta susceptibilidad de las líneas ER+ podría deberse, en parte, a la presencia de un polimorfismo relacionado con el requerimiento de altas concentraciones de fármaco. Finalmente, nuestros resultados sugieren que la administración de RP podría estar desregulando el metabolismo de la valina y la isoleucina.

1.12 Limitaciones del estudio

Nuestro estudio presenta algunas limitaciones. El FBA proporciona un valor de crecimiento óptimo pero múltiples combinaciones de flujos pueden dar lugar a ese valor óptimo, es decir, no existe una solución única, dificultando la comparación entre los flujos de las distintas rutas metabólicas. En este trabajo, esa dificultad se resolvió usando técnicas de remuestreo. Sin embargo, es todavía necesaria una mejora en los procesos computacionales para hacerlos más eficientes. Con respecto a los experimentos de proteómica, a pesar de que mejoran la exactitud de las predicciones, debido a que, a diferencia de la expresión génica, proporcionan mediciones directas de los niveles de enzima disponibles, en este momento esta aproximación sólo proporciona valores para el 57% de las reacciones incluidas en la Recon2 que tienen GPR. La expresión génica, sin embargo, con la limitación de ser una medición indirecta de la abundancia enzimática, proporciona un cuadro más completo. Es interesante reseñar que el FBA no refleja los cambios provocados por la administración de RP. A pesar del potencial de este método, sólo tiene en cuenta diferencias a nivel metabólico. Es bien sabido que la inhibición de mTOR conlleva cambios masivos en la homeostasis celular, por lo que parece razonable que los

modelos metabólicos no predigan estos cambios, ya que sólo tienen en cuenta información relacionada con rutas metabólicas clásicas.

1.13 Novedad del estudio

En este trabajo se propone un flujo de trabajo para estudiar la respuesta a fármacos que afectan al metabolismo usando métodos experimentales y computacionales que permiten proponer nuevas hipótesis y caracterizar la respuesta a nivel molecular, funcional y metabólico, proporcionando una visión global del proceso. Además, se han caracterizado patrones de expresión de proteínas diferenciales entre células control y células tratadas. Asimismo, se ha desarrollado un flujo de trabajo computacional para evaluar el impacto de las alteraciones metabólicas en la tasa de crecimiento tumoral y celular mediante datos de proteómica. Las tasas de crecimiento predichas por nuestro modelo concuerdan con los datos observados experimentalmente. Además, los MGP han demostrado su utilidad para el estudio de los efectos relacionados con procesos biológicos en lugar de considerar los genes o proteínas de manera individual. Nuestra aproximación muestra que estos análisis proporcionan información complementaria y que pueden emplearse para proponer nuevas hipótesis sobre mecanismos de acción y respuesta para posteriormente validarlos experimentalmente. Finalmente, este tipo de análisis podría utilizarse en un futuro para estudiar patrones metabólicos de muestras de tumores con una respuesta diferente a fármacos que tienen como diana procesos metabólicos. Hasta aquí, estos resultados se encuentran publicados en la revista *Oncotarget* (155).

1.14 Porcentaje de coincidencia entre el valor más frecuente del remuestreo y la primera solución que proporciona el FBA

Tras las dificultades encontradas debidas al tiempo de computación que se necesita para llevar a cabo un remuestreo, se decidió estudiar cuál era la variación entre la moda, es decir, el valor más frecuente que toma cada una de las reacciones de la Recon2, y el valor que nos proporciona la función *optimizeCbmodel* de COBRA Toolbox que es más eficiente computacionalmente hablando. Debido a que los porcentajes de coincidencia son muy elevados (Tabla 12) se procedió a considerar a partir de este momento el valor de cada flujo proporcionado por la función *optimizeCbmodel* como una solución representativa. Esto nos permitía ahorrar tiempo de computación.

2. Aplicación del *Flux Balance Analysis* a datos de proteómica provenientes de muestras FFPE de tumores de cáncer de mama

2.1 Tasa de crecimiento tumoral para datos de proteómica de muestras FFPE de tumores de cáncer de mama

Se realizó el FBA para los datos de proteómica provenientes de 96 tumores de cáncer de mama que habían sido previamente caracterizados como *ER-true*, *TN-like* y TNBC. La tasa de crecimiento tumoral predicha para los tumores caracterizados como *ER-true* es significativamente menor que para los *TN-like*. Como era de esperar, no existen diferencias significativas entre los *TN-like* y los TNBC. Estos datos concuerdan con el conocimiento clínico, ya que los TNBC tienen un peor pronóstico por su mayor tasa de crecimiento. Además, estos resultados confirman la hipótesis previa de que dentro de los tumores ER+ existe un grupo de tumores con un comportamiento y pronóstico similar a los TNBC. Estos resultados fueron publicados en la revista *Scientific Reports* (30).

2.2 Predictor de recaída a distancia basado en las actividades de los flujos de las muestras de proteómica

Como ya se ha demostrado anteriormente que la solución aportada por la función *optimizeCbmodel* era una solución representativa, se emplearon los datos de flujo obtenidos con esta función para calcular las actividades de los flujos para los datos de proteómica de 96 pacientes de cáncer de mama y asociarlas con pronóstico.

Mediante las actividades de flujo se establecieron diferencias entre subtipos de cáncer de mama a nivel de detoxificación de ROS y en la ruta de interconversión de nucleótidos. El estrés oxidativo está relacionado con la agresividad tumoral en tumores ER+ (156). Los TNBC también tienen una alta presencia de ROS (157). Como es de esperar, nuestro modelo predice una menor detoxificación de ROS en tumores ER+. La ruta de interconversión de nucleótidos contiene información acerca del metabolismo del ATP y el GTP. No han sido previamente descritas en la literatura diferencias en la producción de ATP y GTP entre los subtipos de cáncer de mama.

Asimismo, es posible asociar estas actividades de flujo con recaída a distancia en esta cohorte. Se encontraron algunas actividades de flujo asociadas con pronóstico y se construyó un predictor. Este predictor consta de la actividad de los flujos de tres rutas metabólicas distintas: la vitamina A, la tetrahidrobiopterina y la beta-alanina. Hasta dónde nosotros sabemos es la primera vez que se asocian datos provenientes de FBA con pronóstico en cáncer. La vitamina A o retinol ya había sido previamente asociada a recaída en cáncer de mama (158) pero no hay

información previa acerca del metabolismo de la tetrahidrobiopterina o la beta-alanina. Además, el valor pronóstico de este predictor se mantiene en los subtipos ER+.

Está ampliamente establecido que los TNBC presentan diferencias en el metabolismo con respecto a los ER+ (30, 41). Por este motivo se construyó un predictor teniendo en cuenta sólo los TNBC. Esta firma contiene las actividades de los flujos de la glucolisis y el metabolismo del glutamato. Sin embargo, el análisis de supervivencia y el análisis multivariante no son significativos, probablemente debido al bajo número de muestras en este grupo. Está descrito un incremento en la captación de glucosa y la producción de lactato en líneas celulares TNBC (41). Está también establecido que la lactato deshidrogenasa B (*LDHB*) y otros genes relacionados con glucolisis se encuentras hiperactivados en tumores TNBC al compararlos con los otros subtipos, así como que la sobreexpresión de *LDHB* está asociada con un peor pronóstico (159). Por otro lado, los tumores TNBC presentan una desregulación de la glutaminolisis y exhiben frecuentemente una expresión más alta de proteínas relacionas con el metabolismo de la glutamina que el resto de subtipos (160, 161).

2.3 Limitaciones del estudio

La principal limitación del estudio es la falta de una validación externa de estos predictores en otra cohorte. Otra limitación es el reducido número de muestras de cada subtipo. Sería necesaria una validación en una cohorte de pacientes más grande. Por otro lado, el número de proteínas detectadas con las actuales mejoras de la técnica ha aumentado, pudiendo aportar una información a nivel de GPR más completa de la que se ha usado en este trabajo.

2.4 Novedad del estudio

Las actividades de los flujos es un método que se propuso para comparar las distribuciones de los flujos entre células control y células tratadas con fármacos que afectan al metabolismo (155). En estos análisis se asocian por primera vez los flujos con el valor pronóstico en una serie de pacientes, demostrando así la utilidad de las actividades de los flujos como variables resumen. Además, se establecen diferencias entre los subtipos de cáncer de mama en la detoxificación de ROS y la interconversión de nucleótidos. Este abordaje podría ser útil a la hora de proponer rutas metabólicas candidatas a ser dianas de tratamientos farmacológicos. Estos resultados se encuentran bajo revisión en la revista *Future Oncology*.

3. Estudio de asociación de datos de metabolómica con los resultados obtenidos en el *Flux Balance Analysis* y con datos de expresión génica en una cohorte de pacientes de cáncer de mama

Los datos de Terunuma *et al.* (112) ya han sido previamente utilizados por el consorcio *The Cancer Genome Atlas* para correlacionar datos de expresión génica con datos de metabolómica mediante coeficientes de Pearson, estableciendo que existía una alta correlación entre ambos tipos de datos (162). En este trabajo, usando estos mismos datos, se emplearon MGP por primera vez en datos de metabolómica provenientes de muestras tumorales y en estos datos de metabolómica en combinación con datos de expresión génica y de actividades de flujo con el objetivo de establecer asociaciones.

3.1 Análisis basados en datos de metabolómica

En primer lugar, se evaluó si era posible asociar los datos de metabolómica con la supervivencia global en pacientes con cáncer de mama. Se generó un predictor relacionado con supervivencia global usando los datos de abundancia de los siguientes metabolitos: glutamina, deoxicarnitina, butirilcarnitina, glicerofosforilcolina y 2-hidroxipalmitato. Los tres primeros ya habían sido previamente asociados con pronóstico en cáncer de mama (163, 164). En el caso del 2-hidroxipalmitato, sin embargo, no hay descrita ninguna relación con supervivencia en cáncer. Además, en el trabajo de Terunuma *et al.* asocian el 2-hidroxiglutarato con peor pronóstico en cáncer de mama (112). El 2-hidroxiglutarato es un intermediario de la glutamina en el ciclo de Krebs, implicado en la conversión de glutamina a lactato, lo que se conoce como glutaminolisis (38). Estos resultados ponen de manifiesto la relevancia del metabolismo de la glutamina en el pronóstico del cáncer de mama, sugiriendo la utilidad de fármacos que tengan como diana esta ruta metabólica, como por ejemplo, la gamma-L-glutamil-p-nitroanilida, que ya se ha descrito que afecta a la proliferación en células de cáncer de pulmón (165). Además, el signo negativo de la glutamina en el predictor indica un efecto protector (mayor cantidad de glutamina está asociada con un mejor pronóstico). Esto podría deberse a que la presencia de glutamina indica que ésta no ha sido introducida al ciclo de Krebs y transformada en aminoácidos y lactato, algo que está relacionado con un fenotipo más agresivo y peor pronóstico.

Se construyó una red a partir de los datos de metabolómica y se caracterizó según rutas metabólicas mediante IMPaLA (120). En trabajos previos se había demostrado la utilidad de los MGP para caracterizar funcionalmente redes de genes o proteínas (30, 41, 76, 155), sin embargo, nunca se había utilizado este método para analizar datos de metabolómica. Al igual que ocurre con los genes o las proteínas, los metabolitos se agrupaban siguiendo una coherencia

basada en la información de las rutas metabólicas. Además, es posible caracterizar diferencias entre las rutas metabólicas empleando la actividad de los nodos, encontrando diferencias entre los tumores ER+ y ER-. Según las actividades de los nodos, el nodo de metabolismo de lípidos presenta una mayor actividad en los tumores ER- que en los ER+. Aunque no hay una relación descrita entre las mediciones de lípidos y los subtipos de cáncer de mama, sí está descrito que los tumores ER- presentan una sobreexpresión de genes relacionados con el metabolismo de lípidos con respecto a los ER+ (166). No se ha encontrado sin embargo ninguna relación previamente descrita entre metabolismo de purinas y cáncer de mama. Además, la actividad del nodo de metabolismo de lípidos presenta valor pronóstico.

3.2 Análisis combinado de metabolitos y genes

La red combinada de metabolitos y genes agrupaba a la mayoría de los metabolitos en un mismo nodo. Sin embargo, algunos metabolitos se incluían en nodos formados por genes, a los cuales se les asignó una función mayoritaria mediante análisis ontológico. Al revisar la bibliografía, cuatro de los 20 metabolitos con una identificación positiva estaban relacionados con la función de su nodo respectivo (Tabla 22).

El succinato se encuentra localizado en el nodo de respuesta inmune. Está descrito que el succinato actúa como una señal inflamatoria, induciendo la producción de la citoquina IL-1β a través de HIF1 (124). Además, el succinato aumenta la capacidad de las células dendríticas para actuar como células presentadoras de antígenos, induciendo por tanto una respuesta inmune adaptativa (125).

La citidina también se encuentra localizada en el nodo de respuesta inmune. En el trabajo de Wachowska *et al.* se describe que la administración de 5-aza-2'-deoxi-citidina modula los niveles de MHC e induce al antígeno P1A en células tumorales y modelos de ratón de cáncer de colon y carcinoma epitelial de mama y que, en combinación con terapia fotodinámica, tiene actividad inmunomoduladora (126)

La histamina aparece relacionada en la red con el proceso de angiogénesis. Esta descrito en la bibliografía que la histamina promueve la angiogénesis a través del factor de crecimiento vascular epitelial (VEGF) (127).

Por último, el 1,2-propanediol también aparece en el nodo de angiogénesis. Su derivado sulfoquinovosilacilpropanediol inhibe la angiogénesis en ratones trasplantados de carcinoma pulmonar (128).

Los dieciséis metabolitos restantes requieren un estudio en profundidad con el fin de establecer asociaciones con las funciones de sus respectivos nodos. Estos resultados avalan el potencial de los MGP como herramienta para generar hipótesis sin necesidad de conocimiento previo.

3.3 Análisis combinado de metabolitos y actividades de flujo

Aunque no existen diferencias significativas en la tasa de crecimiento tumoral entre los tumores ER+ y ER-, sí existen diferencias a nivel de las actividades de los flujos. Además, algunas de estas actividades se usaron para construir un predictor de supervivencia. Es interesante destacar que una de estas actividades de flujo es el metabolismo de la glutamina, resultado concordante con el predictor construido a partir de metabolitos que define la glutamina como un metabolito capaz de predecir supervivencia.

Con el objetivo de relacionar los datos de metabolómica con los datos de las actividades de los flujos se combinaron ambos en una nueva red. A diferencia de la red de genes y metabolitos, las actividades de los flujos aparecían distribuidas por toda la red. Es interesante señalar que aparecían como nodos terminales, lo que puede ser debido a que son un resumen final de cada una de las rutas.

IMPaLA nos permite asignar a cada una de las ramas formadas por metabolitos una ruta metabólica mayoritaria y hacer una comparación de cuántos metabolitos están relacionados con la ruta de la actividad de flujo (120). En la mayoría de los casos existía una correspondencia entre los metabolitos del nodo y la actividad de flujo de ese mismo nodo. Este hecho valida el FBA y las actividades de los flujos, basados ambos en expresión génica, como un método para estudiar el metabolismo. Este resultado es, no obstante, muy preliminar, siendo necesario establecer en un futuro un sistema parecido a DAVID que pondere la significación de esa asociación. La asignación de categorías ontológicas en DAVID está corregida por el número de genes que componen esa ontología. IMPaLA sólo proporciona la información de a qué ruta pertenece cada metabolito.

3.4 Limitaciones del estudio

Nuestro estudio es muy preliminar, siendo necesaria la validación en otras cohortes. Además, la metabolómica ha experimentado un extraordinario avance en los últimos años. Debido a la antigüedad de la serie, el número de metabolitos detectados e identificados es menor del que puede conseguirse en la actualidad. Además, los resultados son difícilmente trasladables a la

práctica clínica en este momento debido a que los tumores de esta serie no recogen información sobre el estatus de Her2.

3.5 Novedad del estudio

La metabolómica se postula como una técnica en auge para la búsqueda de biomarcadores en cáncer. A pesar de que los MGP habían demostrado utilidad en el estudio de datos de proteómica y de expresión génica, nunca se había probado su efectividad en datos de metabolómica. En este trabajo se desarrolla un flujo de trabajo para el análisis de este tipo de datos basado en MGP y el análisis de rutas metabólicas mediante IMPaLA, que permite la caracterización de los subtipos de tumores a nivel global de ruta metabólica en lugar de metabolito a metabolito. Además, es posible asociar los datos de metabolómica con parámetros clínicos. Estos resultados se encuentran en revisión en *PLoS One*.

4. Interfaz FLUX en GUIDE para facilitar la realización del *Flux Balance Analysis*

Con la GUIDE de MATLAB es posible crear interfaces de usuario en las que no sea necesario el uso de programación. Con el fin de hacer el análisis del FBA más accesible se creó una interfaz que permitía mediante botones realizar un FBA completo, así como un análisis FVA y de *knockouts*. Ya que COBRA Toolbox es un proyecto con cooperación abierta (https:// opencobra.github.io/) quizá sería interesante poner esta interfaz al servicio de los usuarios.

CONCLUSIONES

CONCLUSIONES

1. Existe una respuesta diferencial de las líneas celulares de cáncer de mama al tratamiento con fármacos que afectan al metabolismo, concretamente a metformina y a rapamicina. Estas diferencias pueden caracterizarse a nivel funcional y de rutas metabólicas mediante el análisis por modelos gráficos probabilísticos y el *Flux Balance Analysis*.

2. El estudio de genotipado de SNP nos ha permitido identificar un polimorfismo candidato a explicar la sensibilidad de las células MDAMB468 a MTF. Esta información acerca de las causas genéticas complementa la información proporcionada por la proteómica.

3. La proteómica es una técnica de alto rendimiento que permite obtener información directa sobre los procesos biológicos. Su principal limitación era la imposibilidad de poder medir una gran cantidad de proteínas. Con la cantidad de proteínas identificadas en este trabajo (7.267 proteínas medidas), esta limitación desaparece, pudiéndose medir casi la totalidad del proteoma celular.

4. El análisis de patrones de expresión diferencial de proteínas sugiere alteraciones en el ciclo celular provocadas por el tratamiento con metformina y rapamicina. Estas alteraciones se confirmaron mediante experimentos de citometría de flujo.

5. Los modelos gráficos probabilísticos han demostrado su utilidad en trabajos previos para la caracterización de las diferencias entre tumores a nivel de procesos biológicos basándonos en proteómica y datos de expresión génica. En este trabajo se va un paso más allá al aplicar estos modelos a datos de metabolómica y de células tratadas con fármacos, demostrando que se mantiene la estructura funcional que confiere su valor a este tipo de modelos, y que es posible la combinación de diferentes tipos de información manteniendo la coherencia de la red.

6. El *Flux Balance Analysis*, un método ampliamente establecido en biotecnología, ha demostrado su utilidad a la hora de modelar alteraciones en cáncer, permitiéndonos estudiar variaciones en el crecimiento tumoral y alteraciones debidas al tratamiento. También se ha podido comprobar la exactitud de sus predicciones mediante *dynamic FBA*.

7. El *Flux Balance Analysis* nos ha permitido definir una activación de enzimas relacionadas con estrés oxidativo en células de cáncer de mama tratadas con MTF que ha podido ser validada experimentalmente.

8. En este trabajo se propone un método para resumir patrones de flujo en lo que hemos llamado las actividades de los flujos. Estas actividades de los flujos han permitido asociar los da-

tos provenientes del *Flux Balance Analysis* con pronóstico y respuesta a fármacos, además de permitirnos establecer rutas metabólicas diferenciales entre grupos de tumores. Asimismo, han permitido asociar datos de metabolómica con resultados del *Flux Balance Analysis*, que están basados en datos de expresión génica.

9. La principal limitación del *Flux Balance Analysis* es la existencia de múltiples soluciones. En este trabajo se demuestra que esto no sesga el análisis, siendo la primera solución ofrecida por los métodos de computación clásicos representativa de la mayoría de las soluciones.

10. La creación de una interfaz para llevar a cabo el *Flux Balance Analysis* permite hacer más accesible este tipo de análisis. Para calcular un *Flux Balance Analysis* con COBRA Toolbox son necesarios conocimientos básicos de programación en MATLAB. Sin embargo, con la interfaz FLUX es posible calcular un *Flux Balance Analysis* de manera más sencilla.

BIBLIOGRAFÍA

BIBLIOGRAFÍA

1. McPherson K, Steel C, Dixon J. Breast cancer-epidemiology, risks, factors and genetics. BMJ; 2000.

2. Siegel R, Naishadham D, Jemal A. Cancer stadistics, 2015. Cancer J Clin; 2015. p. 5-29.

3. Sinn HP, Kreipe H. A Brief Overview of the WHO Classification of Breast Tumors, 4th Edition, Focusing on Issues and Updates from the 3rd Edition. Breast Care (Basel). 2013;8(2):149-54.

4. Li CI, Anderson BO, Daling JR, Moe RE. Trends in incidence rates of invasive lobular and ductal breast carcinoma. JAMA. 2003;289(11):1421-4.

5. Ellis I, Galea M, Broughton N, Locker A, Blamey R, Elston C. Pathological prognostic factors in breast cancer.II.Histological type.Relationship with survival in a large study with long-term follow. Histopathology1992. p. 479-89.

6. Goldhirsch A, Ingle JN, Gelber RD, Coates AS, Thürlimann B, Senn HJ, et al. Thresholds for therapies: highlights of the St Gallen International Expert Consensus on the primary therapy of early breast cancer 2009. Ann Oncol. 2009;20(8):1319-29.

7. Mook S, Schmidt MK, Rutgers EJ, van de Velde AO, Visser O, Rutgers SM, et al. Calibration and discriminatory accuracy of prognosis calculation for breast cancer with the online Adjuvant! program: a hospital-based retrospective cohort study. Lancet Oncol. 2009;10(11):1070-6.

8. Perou CM, Sorlie T, Eisen MB, van de Rijn M, Jeffrey SS, Rees CA, et al. Molecular portraits of human breast tumours. Nature. 2000;406(6797):747-52.

9. Sørlie T, Perou CM, Tibshirani R, Aas T, Geisler S, Johnsen H, et al. Gene expression patterns of breast carcinomas distinguish tumor subclasses with clinical implications. Proc Natl Acad Sci U S A. 2001;98(19):10869-74.

10. Hu Z, Fan C, Oh DS, Marron JS, He X, Qaqish BF, et al. The molecular portraits of breast tumors are conserved across microarray platforms. BMC Genomics. 2006;7:96.

11. Kuiper GG, Enmark E, Pelto-Huikko M, Nilsson S, Gustafsson JA. Cloning of a novel receptor expressed in rat prostate and ovary. Proc Natl Acad Sci U S A. 1996;93(12):5925-30.

12. Fisher B, Costantino J, Redmond C, Poisson R, Bowman D, Couture J, et al. A randomized clinical trial evaluating tamoxifen in the treatment of patients with node-negative breast cancer who have estrogen-receptor-positive tumors. N Engl J Med. 1989;320(8):479-84.

13. Gershanovich M, Chaudri HA, Campos D, Lurie H, Bonaventura A, Jeffrey M, et al. Letrozole, a new oral aromatase inhibitor: randomised trial comparing 2.5 mg daily, 0.5 mg daily and aminoglutethimide in postmenopausal women with advanced breast cancer. Letrozole International Trial Group (AR/BC3). Ann Oncol. 1998;9(6):639-45.

14. Buzdar A, Jonat W, Howell A, Jones SE, Blomqvist C, Vogel CL, et al. Anastrozole, a potent and selective aromatase inhibitor, versus megestrol acetate in postmenopausal women with advanced breast cancer: results of overview analysis of two phase III trials. Arimidex Study Group. J Clin Oncol. 1996;14(7):2000-11.

15. Bonneterre J, Thürlimann B, Robertson JF, Krzakowski M, Mauriac L, Koralewski P, et al. Anastrozole versus tamoxifen as first-line therapy for advanced breast cancer in 668 postmenopausal women: results of the Tamoxifen or Arimidex Randomized Group Efficacy and Tolerability study. J Clin Oncol. 2000;18(22):3748-57.

16. Nabholtz JM, Buzdar A, Pollak M, Harwin W, Burton G, Mangalik A, et al. Anastrozole is superior to tamoxifen as first-line therapy for advanced breast cancer in postmenopausal women: results of a North American multicenter randomized trial. Arimidex Study Group. J Clin Oncol. 2000;18(22):3758-67.

17. Wakeling AE, Dukes M, Bowler J. A potent specific pure antiestrogen with clinical potential. Cancer Res. 1991;51(15):3867-73.

18. Robertson JF, Lindemann JP, Llombart-Cussac A, Rolski J, Feltl D, Dewar J, et al. Fulvestrant 500 mg versus anastrozole 1 mg for the first-line treatment of advanced breast cancer: follow-up analysis from the randomized 'FIRST' study. Breast Cancer Res Treat. 2012;136(2):503-11.

19. Ellis MJ, Llombart-Cussac A, Feltl D, Dewar JA, Jasiówka M, Hewson N, et al. Fulvestrant 500 mg Versus Anastrozole 1 mg for the First-Line Treatment of Advanced Breast Cancer: Overall Survival Analysis From the Phase II FIRST Study. J Clin Oncol. 2015;33(32):3781-7.

20. Slamon DJ, Clark GM, Wong SG, Levin WJ, Ullrich A, McGuire WL. Human breast cancer: correlation of relapse and survival with amplification of the HER-2/neu oncogene. Science. 1987;235(4785):177-82.

21. Wolff AC, Hammond ME, Schwartz JN, Hagerty KL, Allred DC, Cote RJ, et al. American Society of Clinical Oncology/College of American Pathologists guideline recommendations for human epidermal growth factor receptor 2 testing in breast cancer. Arch Pathol Lab Med. 2007;131(1):18-43.

22. Yarden Y, Sliwkowski MX. Untangling the ErbB signalling network. Nat Rev Mol Cell Biol. 2001;2(2):127-37.

23. Slamon DJ, Leyland-Jones B, Shak S, Fuchs H, Paton V, Bajamonde A, et al. Use of chemotherapy plus a monoclonal antibody against HER2 for metastatic breast cancer that overexpresses HER2. N Engl J Med. 2001;344(11):783-92.

24. Valabrega G, Montemurro F, Aglietta M. Trastuzumab: mechanism of action, resistance and future perspectives in HER2-overexpressing breast cancer. Ann Oncol. 2007;18(6):977-84.

25. Foulkes WD, Smith IE, Reis-Filho JS. Triple-negative breast cancer. N Engl J Med. 2010;363(20):1938-48.

26. Lehmann BD, Jovanović B, Chen X, Estrada MV, Johnson KN, Shyr Y, et al. Refinement of Triple-Negative Breast Cancer Molecular Subtypes: Implications for Neoadjuvant Chemotherapy Selection. PLoS One. 2016;11(6):e0157368.

27. Dent R, Trudeau M, Pritchard KI, Hanna WM, Kahn HK, Sawka CA, et al. Triple-negative breast cancer: clinical features and patterns of recurrence. Clin Cancer Res. 2007;13(15 Pt 1):4429-34.

28. De Giorgi U, Rosti G, Frassineti L, Kopf B, Giovannini N, Zumaglini F, et al. High-dose chemotherapy for triple negative breast cancer. Ann Oncol. 18. England2007. p. 202-3.

29. Liedtke C, Mazouni C, Hess KR, Andre F, Tordai A, Mejia JA, et al. Response to neoadjuvant therapy and long-term survival in patients with triple-negative breast cancer. J Clin Oncol. 2008;26(8):1275-81.

30. Gámez-Pozo A, Trilla-Fuertes L, Berges-Soria J, Selevsek N, López-Vacas R, Díaz-Almirón M, et al. Functional proteomics outlines the complexity of breast cancer molecular subtypes. Scientific Reports. 2017;7(1):10100.

31. Untch M, Konecny GE, Paepke S, von Minckwitz G. Current and future role of neoadjuvant therapy for breast cancer. Breast. 2014;23(5):526-37.

32. Hanahan D, Weinberg RA. Hallmarks of cancer: the next generation. Cell. 2011;144(5):646-74.

33. Warburg O. The metabolism of carcinoma cells. J Cancer Res. 1925;9:148-63.

34. Warburg O. On the origin of cancer cells. Science. 1956;123.

35. Moreno-Sanchez R, Rodriguez-Enriquez S, Marin-Hernandez A, Saavedra E. Energy metabolism in tumor cells. Febs j. 2007;274(6):1393-418.

36. Elstrom RL, Bauer DE, Buzzai M, Karnauskas R, Harris MH, Plas DR, et al. Akt stimulates aerobic glycolysis in cancer cells. Cancer Res. 2004;64(11):3892-9.

37. Lum JJ, Bui T, Gruber M, Gordan JD, DeBerardinis RJ, Covello KL, et al. The transcription factor HIF-1alpha plays a critical role in the growth factor-dependent regulation of both aerobic and anaerobic glycolysis. Genes Dev. 2007;21(9):1037-49.

38. DeBerardinis RJ, Mancuso A, Daikhin E, Nissim I, Yudkoff M, Wehrli S, et al. Beyond aerobic glycolysis: transformed cells can engage in glutamine metabolism that exceeds the requirement for protein and nucleotide synthesis. Proc Natl Acad Sci U S A. 2007;104(49):19345-50.

39. Eagle H, Oyama VI, LEVY M, Horton CL, Fleischman R. The growth response of mammalian cells in tissue culture to L-glutamine and L-glutamic acid. J Biol Chem. 1956;218(2):607-16.

40. Wise DR, DeBerardinis RJ, Mancuso A, Sayed N, Zhang XY, Pfeiffer HK, et al. Myc regulates a transcriptional program that stimulates mitochondrial glutaminolysis and leads to glutamine addiction. Proc Natl Acad Sci U S A. 2008;105(48):18782-7.

41. Gámez-Pozo A, Berges-Soria J, Arevalillo JM, Nanni P, López-Vacas R, Navarro H, et al. Combined label-free quantitative proteomics and microRNA expression analysis of breast cancer unravel molecular differences with clinical implications. Cancer Res; 2015. p. 2243-53.

42. Jones NP, Schulze A. Targeting cancer metabolism--aiming at a tumour's sweet-spot. Drug Discov Today. 2012;17(5-6):232-41.

43. Sborov DW, Haverkos BM, Harris PJ. Investigational cancer drugs targeting cell metabolism in clinical development. Expert Opin Investig Drugs. 2015;24(1):79-94.

44. Witters LA. The blooming of the French lilac. J Clin Invest. 2001;108(8):1105-7.

45. Zhou G, Myers R, Li Y, Chen Y, Shen X, Fenyk-Melody J, et al. Role of AMP-activated protein kinase in mechanism of metformin action. J Clin Invest. 2001;108(8):1167-74.

46. El-Mir MY, Nogueira V, Fontaine E, Avéret N, Rigoulet M, Leverve X. Dimethylbiguanide inhibits cell respiration via an indirect effect targeted on the respiratory chain complex I. J Biol Chem. 2000;275(1):223-8.

47. Zakikhani M, Dowling R, Fantus IG, Sonenberg N, Pollak M. Metformin is an AMP kinase-dependent growth inhibitor for breast cancer cells. Cancer Res. 2006;66(21):10269-73.

48. Ellis MJ, Perou CM. The genomic landscape of breast cancer as a therapeutic roadmap. Cancer Discov. 2013;3(1):27-34.

49. Lee JJ, Loh K, Yap YS. PI3K/Akt/mTOR inhibitors in breast cancer. Cancer Biol Med. 2015;12(4):342-54.

50. Beck JT. Potential role for mammalian target of rapamycin inhibitors as first-line therapy in hormone receptor-positive advanced breast cancer. Onco Targets Ther. 2015;8:3629-38.

51. Tengku Din TA, Seeni A, Khairi WN, Shamsuddin S, Jaafar H. Effects of rapamycin on cell apoptosis in MCF-7 human breast cancer cells. Asian Pac J Cancer Prev. 2014;15(24):10659-63.

52. Düvel K, Yecies JL, Menon S, Raman P, Lipovsky AI, Souza AL, et al. Activation of a metabolic gene regulatory network downstream of mTOR complex 1. Mol Cell. 2010;39(2):171-83.

53. Chou TC. Theoretical basis, experimental design, and computerized simulation of synergism and antagonism in drug combination studies. Pharmacol Rev. 2006;58(3):621-81.

54. Chou TC. Drug combination studies and their synergy quantification using the Chou-Talalay method. Cancer Res. 2010;70(2):440-6.

55. Chou T, Martin N. CompuSyn for drug combinations and from general dose-effect analysis. User´s Guide. A computer program for quantitation of synergism and antagonism in drug combinations and the determination of IC_{50}, ED_{50} and LD_{50} values. ComboSyn, Inc Paramus, NJ.; 2007.

56. Nelander S, Wang W, Nilsson B, She QB, Pratilas C, Rosen N, et al. Models from experiments: combinatorial drug perturbations of cancer cells. Mol Syst Biol. 2008;4:216.

57. Aksenov SV, Church B, Dhiman A, Georgieva A, Sarangapani R, Helmlinger G, et al. An integrated approach for inference and mechanistic modeling for advancing drug development. FEBS Lett. 2005;579(8):1878-83.

58. Xiao Y, Gong Y, Lv Y, Lan Y, Hu J, Li F, et al. Gene Perturbation Atlas (GPA): a single-gene perturbation repository for characterizing functional mechanisms of coding and non-coding genes. Sci Rep. 2015;5:10889.

59. Duan Q, Wang Z, Fernandez NF, Rouillard AD, Tan CM, Benes CH, et al. Drug/Cell-line Browser: interactive canvas visualization of cancer drug/cell-line viability assay datasets. Bioinformatics. 2014;30(22):3289-90.

60. Wilkins MR, Sanchez JC, Gooley AA, Appel RD, Humphery-Smith I, Hochstrasser DF, et al. Progress with proteome projects: why all proteins expressed by a genome should be identified and how to do it. Biotechnol Genet Eng Rev. 1996;13:19-50.

61. Ezkurdia I, Juan D, Rodriguez JM, Frankish A, Diekhans M, Harrow J, et al. Multiple evidence strands suggest that there may be as few as 19,000 human protein-coding genes. Hum Mol Genet. 2014;23(22):5866-78.

62. Domon B, Aebersold R. Mass spectrometry and protein analysis. Science. 2006;312(5771):212-7.

63. Meier F, Geyer PE, Virreira Winter S, Cox J, Mann M. BoxCar acquisition method enables single-shot proteomics at a depth of 10,000 proteins in 100 minutes. Nat Methods. 2018.

64. Cox J, Mann M. MaxQuant enables high peptide identification rates, individualized p.p.b.-range mass accuracies and proteome-wide protein quantification. Nat Biotechnol. 2008;26(12):1367-72.

65. Fiehn O. Metabolomics--the link between genotypes and phenotypes. Plant Mol Biol. 2002;48(1-2):155-71.

66. Oliver SG, Winson MK, Kell DB, Baganz F. Systematic functional analysis of the yeast genome. Trends Biotechnol. 1998;16(9):373-8.

67. Fiehn O, Kopka J, Dörmann P, Altmann T, Trethewey RN, Willmitzer L. Metabolite profiling for plant functional genomics. Nat Biotechnol. 2000;18(11):1157-61.

68. Patti GJ, Yanes O, Siuzdak G. Innovation: Metabolomics: the apogee of the omics trilogy. Nat Rev Mol Cell Biol. 2012;13(4):263-9.

69. Wishart DS. Emerging applications of metabolomics in drug discovery and precision medicine. Nat Rev Drug Discov. 2016;15(7):473-84.

70. Kaushik AK, DeBerardinis RJ. Applications of metabolomics to study cancer metabolism. Biochim Biophys Acta Rev Cancer. 2018;1870(1):2-14.

71. Emwas AH. The strengths and weaknesses of NMR spectroscopy and mass spectrometry with particular focus on metabolomics research. Methods Mol Biol. 2015;1277:161-93.

72. Dunn WB, Broadhurst D, Begley P, Zelena E, Francis-McIntyre S, Anderson N, et al. Procedures for large-scale metabolic profiling of serum and plasma using gas chromatography and liquid chromatography coupled to mass spectrometry. Nat Protoc. 2011;6(7):1060-83.

73. Fuhrer T, Zamboni N. High-throughput discovery metabolomics. Curr Opin Biotechnol. 2015;31:73-8.

74. Schwarz G. Estimating the dimension of a model. Ann Stat1978. p. 461-4.

75. Lauritzen S. Graphical Models. Oxford,UK.: Oxford University Press1996.

76. de Velasco G, Trilla-Fuertes L, Gamez-Pozo A, Urbanowicz M, Ruiz-Ares G, Sepúlveda JM, et al. Urothelial cancer proteomics provides both prognostic and functional information. Sci Rep. 2017;7(1):15819.

77. Zapater-Moros A, Gámez-Pozo A, Prado-Vázquez G, Trilla-Fuertes L, Arevalillo JM, Díaz-Almirón M, et al. Probabilistic graphical models relate immune status with response to neoadjuvant chemotherapy in breast cancer. Oncotarget. 2018;9(45):27586-94.

78. Varma A, Palsson BO. Parametric sensitivity of stoichiometric flux balance models applied to wild-type Escherichia coli metabolism. Biotechnol Bioeng. 1995;45(1):69-79.

79. Pramanik J, Keasling JD. Stoichiometric model of Escherichia coli metabolism: incorporation of growth-rate dependent biomass composition and mechanistic energy requirements. Biotechnol Bioeng. 1997;56(4):398-421.

80. Edwards J. Functional genomics and the computational analysis of bacterial metabolism. San Diego: University of California; 1999.

81. Orth J, Thiele I, Palsson B. What is flux balance analysis? : Nat Biotechnol; 2010. p. 245-8.

82. Edwards J, Ibarra R, Palsson B. In silico predictions of Eschecherichia coli metabolic capabilities are consistent with experimental data. Nat Biotechnol; 2001. p. 125-30.

83. Thiele I, Swainston N, Fleming RM, Hoppe A, Sahoo S, Aurich MK, et al. A community-driven global reconstruction of human metabolism. Nat Biotechnol. 2013;31(5):419-25.

84. Schilling C, Palsson B. The underlying pathway structure of biochemical reaction networks. Proc Natl Acad Sci USA; 1998. p. 4193-8.

85. Asgari Y, Zabihinpour Z, Salehzadeh A, Schreiber F, Masoudi-Nejad A. Alterations in cancer cell metabolism: The Warburg effect and metabolic adaptation. Genomics; 2015. p. 275-81.

86. Brunk E, Sahoo S, Zielinski DC, Altunkaya A, Dräger A, Mih N, et al. Recon3D enables a three-dimensional view of gene variation in human metabolism. Nat Biotechnol. 2018;36(3):272-81.

87. Palsson B. Systems Biology: Properties of reconstructde networks. Cambridge University Press; 2007.

88. Schellenberger J, Que R, Fleming R, Thiele I, Orth J, Feist A, et al. Quantitative prediction of cellular metabolism with constraint-based models: the COBRA Toolbox v2.0. Nature Protocols; 2011. p. 1290-307.

89. Thiele I, Palsson B. A protocol for generating a high-quality genome-scale metabolic reconstruction. Nat Protoc; 2010. p. 93-121.

90. Díaz Almirón M. Desarrollo del algoritmo CAPRI para la estimación de la abundancia de complejos enzimáticos y su aplicación en la modelización del metabolismo en cáncer de mama. Universidad Autónoma de Madrid; 2017.

91. Machado D, Herrgård M. Systematic evaluation of methods for integration of transcriptomic data into constraint-based models of metabolism. PLoS Comput Biol. 2014;10(4):e1003580.

92. Akesson M, Förster J, Nielsen J. Integration of gene expression data into genome-scale metabolic models. Metab Eng. 2004;6(4):285-93.

93. Becker SA, Palsson BO. Context-specific metabolic networks are consistent with experiments. PLoS Comput Biol. 2008;4(5):e1000082.

94. Zur H, Ruppin E, Shlomi T. iMAT: an integrative metabolic analysis tool. Bioinformatics. 2010;26(24):3140-2.

95. Brandes A, Lun D, Ip K, Zucker J, Colijn C, Weiner B, et al. Inferring carbon sources from gene expression profiles using metabolic flux models. PLOS One; 2012. p. e36947.

96. Colijn C, Brandes A, Zucker J, Lun D, Weiner B, Farhat M, et al. Interpreting expression data with metabolic flux models: Predicting Mycobacterium tuberculosis mycolic acid production. PLOS Comput Bio; 2009.

97. Lee D, Smallbone K, Dunn WB, Murabito E, Winder CL, Kell DB, et al. Improving metabolic flux predictions using absolute gene expression data. BMC Syst Biol. 2012;6:73.

98. Jerby L, Wolf L, Denkert C, Stein G, Hilvo M, Oresic M, et al. Metabolic associations of reduced proliferation oxidative stress in advanced breast cancer. Cancer Res; 2012. p. 12-20.

99. Song HS, Reifman J, Wallqvist A. Prediction of metabolic flux distribution from gene expression data based on the flux minimization principle. PLoS One. 2014;9(11):e112524.

100. Edwards JS, Palsson BO. The Escherichia coli MG1655 in silico metabolic genotype: its definition, characteristics, and capabilities. Proc Natl Acad Sci U S A. 2000;97(10):5528-33.

101. Förster J, Famili I, Palsson B, Nielsen J. Large-scale evaluation of in silico gene deletions in Sacharomyces cerevisae. OMICS; 2003. p. 193-202.

102. Yizhak K, Gaude E, Le Dévédec S, Waldman Y, Stein G, van de Water B, et al. Phenotype-based cell-specific metabolic modeling reveals metabolic liabilities of cancer. E-life; 2014.

103. Gatto F, Miess H, Schulze A, Nielsen J. Flux balance analysis predicts essential genes in clear cell renal cell carcinoma metabolism. Sci Rep. 2015;5:10738.

104. Varma A, Palsson BO. Stoichiometric flux balance models quantitatively predict growth and metabolic by-product secretion in wild-type Escherichia coli W3110. Appl Environ Microbiol. 1994;60(10):3724-31.

105. Resendis-Antonio O, Checa A, Encarnación S. Modeling core metabolism in cancer cells: Surveying the topology underling the Warburg effect. PLOS One; 2010. p. e12383.

106. Schellenberger J. Monte Carlo simulation in Systems Biology. Bioinformatics and Systems Biology. San Diego. University of California; 2010. p. 162.

107. Vázquez A, Liu J, Zhou Y, Oltvai Z. Catabolic efficiency of aerobic glycolysis: the Warburg effect revisited. BMC Syst Biol; 2010. p. doi: 10.1186/752-0509.

108. Shlomi T, Benyamini T, Gottlieb E, Sharan R, Ruppin E. Genome-scale metabolic modeling elucidates the role of proliferative adaptation in causing the Warbug effect. PLOS Comput Biol; 2011. p. e1002018.

109. Folger O, Jerby L, Frezza C, Gottlieb E, Ruppin E, Shlomi T. Predicting selective drug targets in cancer through metabolic networks. Mol Syst Biol; 2011. p. 501.

110. Agren R, Bordel S, Mardinoglu A, Pornputtapong N, Nookaew I, Nielsen J. Reconstruction of genome-scale active metabolic networks for 69 human cell types and 16 cancer types using INIT. PLOS Comput Biol; 2012. p. e1002518.

111. Wang Y, Eddy J,Price N. Reconstruction of genome-scale metabolic models for 126 human tissues using mCADRE. BMC Syst Biol; 2012. p. 153. doi: 10.1186/752-0509-6-153.

112. Terunuma A, Putluri N, Mishra P, Mathé EA, Dorsey TH, Yi M, et al. MYC-driven accumulation of 2-hydroxyglutarate is associated with breast cancer prognosis. J Clin Invest. 2014;124(1):398-412.

113. Tyanova S, Temu T, Sinitcyn P, Carlson A, Hein MY, Geiger T, et al. The Perseus computational platform for comprehensive analysis of (prote)omics data. Nat Methods. 2016;13(9):731-40.

114. Olsen JV, de Godoy LM, Li G, Macek B, Mortensen P, Pesch R, et al. Parts per million mass accuracy on an Orbitrap mass spectrometer via lock mass injection into a C-trap. Mol Cell Proteomics. 2005;4(12):2010-21.

115. Cox J, Neuhauser N, Michalski A, Scheltema RA, Olsen JV, Mann M. Andromeda: a peptide search engine integrated into the MaxQuant environment. J Proteome Res. 2011;10(4):1794-805.

116. Huang dW, Sherman BT, Lempicki RA. Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. Nat Protoc. 2009;4(1):44-57.

117. Davis AP, Grondin CJ, Johnson RJ, Sciaky D, King BL, McMorran R, et al. The Comparative Toxicogenomics Database: update 2017. Nucleic Acids Res. 2017;45(D1):D972-D8.

118. Abreu G, Edwards D, Labouriau R. High-Dimensional Graphical Model Search with the gRapHD R Package Journal of Statistical Software2010. p. 1-18.

119. Shannon P, Markiel A, Ozier O, Baliga NS, Wang JT, Ramage D, et al. Cytoscape: a software environment for integrated models of biomolecular interaction networks. Genome Res. 2003;13(11):2498-504.

120. Cavill R, Kamburov A, Ellis JK, Athersuch TJ, Blagrove MS, Herwig R, et al. Consensus-phenotype integration of transcriptomic and metabolomic data implies a role for metabolism in the chemosensitivity of tumour cells. PLoS Comput Biol. 2011;7(3):e1001113.

121. Barker BE, Sadagopan N, Wang Y, Smallbone K, Myers CR, Xi H, et al. A robust and efficient method for estimating enzyme complex abundance and metabolic flux from expression data. Comput Biol Chem. 2015;59 Pt B:98-112.

122. Simon R. Roadmap for developing and validating therapeutically relevant genomic classifiers. J Clin Oncol. 2005;23(29):7332-41.

123. Wang Y, Wei J, Li L, Fan C, Sun Y. Combined Use of Metformin and Everolimus Is Synergistic in the Treatment of Breast Cancer Cells. Oncol Res. 2014;22(4):193-201.

124. Tannahill GM, Curtis AM, Adamik J, Palsson-McDermott EM, McGettrick AF, Goel G, et al. Succinate is an inflammatory signal that induces IL-1β through HIF-1α. Nature. 2013;496(7444):238-42.

125. Jiang S, Yan W. Succinate in the cancer-immune cycle. Cancer Lett. 2017;390:45-7.

126. Wachowska M, Gabrysiak M, Muchowicz A, Bednarek W, Barankiewicz J, Rygiel T, et al. 5-Aza-2'-deoxycytidine potentiates antitumour immune response induced by photodynamic therapy. Eur J Cancer. 2014;50(7):1370-81.

127. Lu Q, Wang C, Pan R, Gao X, Wei Z, Xia Y, et al. Histamine synergistically promotes bFGF-induced angiogenesis by enhancing VEGF production via H1 receptor. J Cell Biochem. 2013;114(5):1009-19.

128. Ruike T, Kanai Y, Iwabata K, Matsumoto Y, Murata H, Ishima M, et al. Distribution and metabolism of 14C-Sulfoquinovosylacylpropanediol (14C-SQAP) after a single intravenous administration in tumor-bearing mice. Xenobiotica. 2018:1-45.

129. Motamedian E, Taheri E, Bagheri F. Proliferation inhibition of cisplatin-resistant ovarian cancer cells using drugs screened by integrating a metabolic model and transcriptomic data. Cell Prolif. 2017;50(6).

130. Sacco F, Silvestri A, Posca D, Pirrò S, Gherardini PF, Castagnoli L, et al. Deep Proteomics of Breast Cancer Cells Reveals that Metformin Rewires Signaling Networks Away from a Pro-growth State. Cell Syst. 2016;2(3):159-71.

131. Buler M, Aatsinki SM, Izzi V, Uusimaa J, Hakkola J. SIRT5 is under the control of PGC-1α and AMPK and is involved in regulation of mitochondrial energy metabolism. FASEB J. 2014;28(7):3225-37.

132. Gomes P, Outeiro TF, Cavadas C. Emerging Role of Sirtuin 2 in the Regulation of Mammalian Metabolism. Trends Pharmacol Sci. 2015;36(11):756-68.

133. Lin WM, Fisher DE. Signaling and Immune Regulation in Melanoma Development and Responses to Therapy. Annu Rev Pathol. 2017;12:75-102.

134. Tang LH, Contractor T, Clausen R, Klimstra DS, Du YC, Allen PJ, et al. Attenuation of the retinoblastoma pathway in pancreatic neuroendocrine tumors due to increased cdk4/cdk6. Clin Cancer Res. 2012;18(17):4612-20.

135. Knudsen ES, Witkiewicz AK. Defining the transcriptional and biological response to CDK4/6 inhibition in relation to ER+/HER2- breast cancer. Oncotarget. 2016;7(43):69111-23.

136. Gonzalez J, Harris T, Childs G, Prystowsky MB. Rapamycin blocks IL-2-driven T cell cycle progression while preserving T cell survival. Blood Cells Mol Dis. 2001;27(3):572-85.

137. Wang XC, Tian J, Tian LL, Wu HL, Meng AM, Ma TH, et al. Role of Cks1 amplification and overexpression in breast cancer. Biochem Biophys Res Commun. 2009;379(4):1107-13.

138. Cui Y, Huang Q, Auman JT, Knight B, Jin X, Blanchard KT, et al. Genomic-derived markers for early detection of calcineurin inhibitor immunosuppressant-mediated nephrotoxicity. Toxicol Sci. 2011;124(1):23-34.

139. Pannu V, Rida PC, Ogden A, Turaga RC, Donthamsetty S, Bowen NJ, et al. HSET overexpression fuels tumor progression via centrosome clustering-independent mechanisms in breast cancer patients. Oncotarget. 2015;6(8):6076-91.

140. Davies TH, Ning YM, Sánchez ER. Differential control of glucocorticoid receptor hormone-binding function by tetratricopeptide repeat (TPR) proteins and the immunosuppressive ligand FK506. Biochemistry. 2005;44(6):2030-8.

141. Vilasco M, Communal L, Mourra N, Courtin A, Forgez P, Gompel A. Glucocorticoid receptor and breast cancer. Breast Cancer Res Treat. 2011;130(1):1-10.

142. He H, Sun Y. Ribosomal protein S27L is a direct p53 target that regulates apoptosis. Oncogene. 2007;26(19):2707-16.

143. Rosenbluth JM, Mays DJ, Jiang A, Shyr Y, Pietenpol JA. Differential regulation of the p73 cistrome by mammalian target of rapamycin reveals transcriptional programs of mesenchymal differentiation and tumorigenesis. Proc Natl Acad Sci U S A. 2011;108(5):2076-81.

144. Belkaid A, Duguay SR, Ouellette RJ, Surette ME. 17β-estradiol induces stearoyl-CoA desaturase-1 expression in estrogen receptor-positive breast cancer cells. BMC Cancer. 2015;15:440.

145. Boulay A, Rudloff J, Ye J, Zumstein-Mecker S, O'Reilly T, Evans DB, et al. Dual inhibition of mTOR and estrogen receptor signaling in vitro induces cell death in models of breast cancer. Clin Cancer Res. 2005;11(14):5319-28.

146. O'Reilly T, McSheehy PM, Wartmann M, Lassota P, Brandt R, Lane HA. Evaluation of the mTOR inhibitor, everolimus, in combination with cytotoxic antitumor agents using human tumor models in vitro and in vivo. Anticancer Drugs. 2011;22(1):58-78.

147. Yee KW, Zeng Z, Konopleva M, Verstovsek S, Ravandi F, Ferrajoli A, et al. Phase I/II study of the mammalian target of rapamycin inhibitor everolimus (RAD001) in patients with relapsed or refractory hematologic malignancies. Clin Cancer Res. 2006;12(17):5165-73.

148. Fingar DC, Richardson CJ, Tee AR, Cheatham L, Tsou C, Blenis J. mTOR controls cell cycle progression through its cell growth effectors S6K1 and 4E-BP1/eukaryotic translation initiation factor 4E. Mol Cell Biol. 2004;24(1):200-16.

149. Smallbone K, Simeonidis E. Flux balance analysis: a geometric perspective. J Theor Biol. 2009;258(2):311-5.

150. Dai J, Liu M, Ai Q, Lin L, Wu K, Deng X, et al. Involvement of catalase in the protective benefits of metformin in mice with oxidative liver injury. Chem Biol Interact. 2014;216:34-42.

151. Kukidome D, Nishikawa T, Sonoda K, Imoto K, Fujisawa K, Yano M, et al. Activation of AMP-activated protein kinase reduces hyperglycemia-induced mitochondrial reactive oxygen species production and promotes mitochondrial biogenesis in human umbilical vein endothelial cells. Diabetes. 2006;55(1):120-7.

152. Volarevic V, Misirkic M, Vucicevic L, Paunovic V, Simovic Markovic B, Stojanovic M, et al. Metformin aggravates immune-mediated liver injury in mice. Arch Toxicol. 2015;89(3):437-50.

153. Vannini F, Kashfi K, Nath N. The dual role of iNOS in cancer. Redox Biol. 2015;6:334-43.

154. Cuyàs E, Corominas-Faja B, Joven J, Menendez JA. Cell cycle regulation by the nutrient-sensing mammalian target of rapamycin (mTOR) pathway. Methods Mol Biol. 2014;1170:113-44.

155. Trilla-Fuertes L, Gámez-Pozo A, Arevalillo JM, Díaz-Almirón M, Prado-Vázquez G, Zapater-Moros A, et al. Molecular characterization of breast cancer cell response to metabolic drugs. Oncotarget. 2018;9(11):9645-60.

156. Mahalingaiah PK, Ponnusamy L, Singh KP. Chronic oxidative stress causes estrogen-independent aggressive phenotype, and epigenetic inactivation of estrogen receptor alpha in MCF-7 breast cancer cells. Breast Cancer Res Treat. 2015;153(1):41-56.

157. Pelicano H, Zhang W, Liu J, Hammoudi N, Dai J, Xu RH, et al. Mitochondrial dysfunction in some triple-negative breast cancer cell lines: role of mTOR pathway and therapeutic potential. Breast Cancer Res. 2014;16(5):434.

158. Formelli F, Meneghini E, Cavadini E, Camerini T, Di Mauro MG, De Palo G, et al. Plasma retinol and prognosis of postmenopausal breast cancer patients. Cancer Epidemiol Biomarkers Prev. 2009;18(1):42-8.

159. McCleland ML, Adler AS, Shang Y, Hunsaker T, Truong T, Peterson D, et al. An integrated genomic screen identifies LDHB as an essential gene for triple-negative breast cancer. Cancer Res. 2012;72(22):5812-23.

160. Kim S, Kim DH, Jung WH, Koo JS. Expression of glutamine metabolism-related proteins according to molecular subtype of breast cancer. Endocr Relat Cancer. 2013;20(3):339-48.

161. Lampa M, Arlt H, He T, Ospina B, Reeves J, Zhang B, et al. Glutaminase is essential for the growth of triple-negative breast cancer cells with a deregulated glutamine metabolism pathway and its suppression synergizes with mTOR inhibition. PLoS One. 2017;12(9):e0185092.

162. Peng X, Chen Z, Farshidfar F, Xu X, Lorenzi PL, Wang Y, et al. Molecular Characterization and Clinical Relevance of Metabolic Expression Subtypes in Human Cancers. Cell Rep. 2018;23(1):255-69.e4.

163. Bhowmik SK, Ramirez-Peña E, Arnold JM, Putluri V, Sphyris N, Michailidis G, et al. EMT-induced metabolite signature identifies poor clinical outcome. Oncotarget. 2015;6(40):42651-60.

164. Cao MD, Sitter B, Bathen TF, Bofin A, Lønning PE, Lundgren S, et al. Predicting long-term survival and treatment response in breast cancer patients receiving neoadjuvant chemotherapy by MR metabolic profiling. NMR Biomed. 2012;25(2):369-78.

165. Hassanein M, Hoeksema MD, Shiota M, Qian J, Harris BK, Chen H, et al. SLC1A5 mediates glutamine transport required for lung cancer cell growth and survival. Clin Cancer Res. 2013;19(3):560-70.

166. Wang J, Shidfar A, Ivancic D, Ranjan M, Liu L, Choi MR, et al. Overexpression of lipid metabolism genes and PBX1 in the contralateral breasts of women with estrogen receptor-negative breast cancer. Int J Cancer. 2017;140(11):2484-97.

ANEXOS

ANEXO 1: METABOLITOS ASOCIADOS A CADA UNA DE LAS ACTIVIDADES DE FLUJO DE CADA RAMA DE LA RED

Rama de la red	Actividad de flujo	Número de metabolitos en esa rama relacionados con la actividad de flujo	Nombre de metabolitos en esa rama relacionados con la actividad de flujo
Rama 1	Glucolisis/gluconeogénesis	3	Citrato, fructosa 6-P, 3-Fosfoglicerato
Rama 1	Síntesis de pirimidinas	2	Guanosina, inosina
Rama 2	Metabolismo de estrógenos y andrógenos	0	
Rama 2	Detoxificación de ROS	0	
Rama 2	Metabolismo de esfingolípidos	0	
Rama 2	Metabolismo de la vitamina A	No cuantificable	
Rama 2	Metabolismo de la vitamina B6	Al menos 1	Glutamato
Rama 2	Metabolismo de la vitamina D	No cuantificable	
Rama 3	Metabolismo de glicina, serina, alanina y treonina	5	Metionina, piruvato, glicina, asparragina, treonina
Rama 3	Metabolismo de la D-alanina	5	metionina, piruvato, glicina, asparragina, treonina
Rama 3	Metabolismo de fructosa y manosa	1	Piruvato
Rama 3	Metabolismo de glutatión	2	Piruvato, glicina
Rama 3	Vía de recuperación de nucleótidos	2	Piruvato, glicina
Rama 3	Metabolismo de fenilalanina	6	Piruvato, valina, isoleucina, tirosina, leucina, histidina
Rama 3	Catabolismo de purinas	2	Piruvato, glicina
Rama 3	Metabolismo de tetrahidrobiopterina	1	Tirosina
Rama 3	Síntesis de triacilglicerol	0	
Rama 3	Metabolismo de vitamina C	No cuantificable	
Rama 3	Interconversión de nucleótidos	2	Piruvato, glicina
Rama 4	Metabolismo de alanina y aspartato	4	Colina, glicerato, citrulina, urea
Rama 4	Metabolismo de butanoato	1	2-hidroxibutirato
Rama 4	Metabolismo de colesterol	0	
Rama 4	Ciclo de Krebs	0	
Rama 4	Catabolismo de coA	1	2-hidroxiestearato
Rama 4	Síntesis de coA	0	
Rama 4	Metabolismo de galactosa	0	
Rama 4	Metabolismo de glutamato	1	Dimetilarginina
Rama 4	Metabolismo de NAD	1	1-metilnicotinamida
Rama 4	Fosforilación oxidativa	1	Citrulina
Rama 4	Ruta de pentosas fosfato	1	Glicerato
Rama 4	Metabolismo de propanoato	2	2-propanediol, 2-hidroxibutirato
Rama 4	Metabolismo de piruvato	1	2-propanediol
Rama 4	Metabolismo de taurina e hipotaurina	0	
Rama 4	Metabolismo de triptófano	0	
Rama 4	Ciclo de la urea	3	Urea, citrulina, creatinina
Rama 4	Metabolismo de beta-alanina	5	6-dihidrouracilo, citrulina, urea, glicerato, colina
Rama 5	Metabolismo de glioxilato y dicarboxilato	1	Fosfato
Rama 5	Metabolismo de vitamina B2		Fosfato
Rama 6	Metabolismo de valina, leucina e isoleucina	0	
Rama 6	Metabolismo de lisina	1	S-adenosilhomocisteína
Rama 6	Metabolismo de histidina	3	S-adenosilhomocisteína, histamina, heme

Rama 6	Metabolismo de arginina y prolina	3	S-adenosilhomocisteína, heme, espermidina
Rama 6	Metabolismo de ácido araquidónico	1	Heme
Rama 7	Metabolismo de aminoazúcares	1	N-acetilneuraminato
Rama 7	Metabolismo de eicosanoides	3	Eicosapentaenoato, araquidonato, glicerol
Rama 7	Metabolismo de folato	1	Glucosa
Rama 7	Metabolismo de glicerofosfolípidos	11	Margarato, oleato, estearato, linolenato, araquidonato, palmitato, linoleato, cis-vacenato, eicosapentaenoato, docosahexaenoato, glicerol
Rama 7	Metabolismo de hialurónico	10	Margarato, oleato, estearato, linolenato, araquidonato, palmitato, linoleato, cis-vacenato, eicosapentaenoato, docosahexaenoato
Rama 7	Metabolismo de metionina y cisteína	0	
Rama 7	Metabolismo de fosfatidilinositol fosfato	1	Glicerol
Rama 7	Catabolismo de pirimidinas	1	Ornitina
Rama 7	Metabolismo de almidón y sacarosa	1	Glucosa
Rama 7	Metabolismo de esteroides	2	Araquidonato, palmitato
Rama 7	Metabolismo de tirosina	1	Succinato

ANEXO 2: CÓDIGO DE LA APLICACIÓN FLUX

```
function varargout = FLUX(varargin)
% FLUX MATLAB code for FLUX.fig
% FLUX, by itself, creates a new FLUX or raises the existing %singleton*.
%
% H = FLUX returns the handle to a new FLUX or the handle to
% the existing singleton*.
%FLUX('CALLBACK',hObject,eventData,handles,...) calls the local
% function named CALLBACK in FLUX.M with the given input arguments.
%
%FLUX('Property','Value',...) creates a new FLUX or raises the
% existing singleton*.  Starting from the left, property value pairs
% are applied to the GUI before FLUX_OpeningFcn gets called.  An
% unrecognized property name or invalid value makes property application
% stop.  All inputs are passed to FLUX_OpeningFcn via varargin.
%
% *See GUI Options on GUIDE's Tools menu.  Choose "GUI allows only one
%instance to run (singleton)".
%
% See also: GUIDE, GUIDATA, GUIHANDLES

% Edit the above text to modify the response to help FLUX

% Last Modified by GUIDE v2.5 28-Sep-2017 12:43:31

% Begin initialization code - DO NOT EDIT
gui_Singleton = 1;
gui_State = struct('gui_Name',       mfilename, ...
                   'gui_Singleton',  gui_Singleton, ...
                   'gui_OpeningFcn', @FLUX_OpeningFcn, ...
                   'gui_OutputFcn',  @FLUX_OutputFcn, ...
                   'gui_LayoutFcn',  [] , ...
                   'gui_Callback',   []);
if nargin && ischar(varargin{1})
    gui_State.gui_Callback = str2func(varargin{1});
end

if nargout
    [varargout{1:nargout}] = gui_mainfcn(gui_State, varargin{:});
else
    gui_mainfcn(gui_State, varargin{:});
end
% End initialization code - DO NOT EDIT

% --- Executes just before FLUX is made visible.
function FLUX_OpeningFcn(hObject, eventdata, handles, varargin)
% This function has no output args, see OutputFcn.
% hObject    handle to figure
% eventdata  reserved - to be defined in a future version of MATLAB
% handles    structure with handles and user data (see GUIDATA)
% varargin   command line arguments to FLUX (see VARARGIN)

% Choose default command line output for FLUX
handles.output = hObject;

% Update handles structure
guidata(hObject, handles);

% UIWAIT makes FLUX wait for user response (see UIRESUME)
% uiwait(handles.figure1);

% --- Outputs from this function are returned to the command line.
function varargout = FLUX_OutputFcn(hObject, eventdata, handles)
% varargout  cell array for returning output args (see VARARGOUT);
% hObject    handle to figure
% eventdata  reserved - to be defined in a future version of MATLAB
```

```
% handles    structure with handles and user data (see GUIDATA)

% Get default command line output from handles structure
varargout{1} = handles.output;
%init Cobra
initCobraToolbox();
savepath();
changeCobraSolver('glpk','LP');

% --- Executes on button press in recon.
function recon_Callback(hObject, eventdata, handles)
% hObject    handle to recon (see GCBO)
% eventdata  reserved - to be defined in a future version of MATLAB
% handles    structure with handles and user data (see GUIDATA)
%load recon
model = readCbModel();
set (handles.text2, 'String','Ready');
assignin ('base', 'model',  model);
% --- Executes on button press in of.
function of_Callback(hObject, eventdata, handles)
% hObject    handle to of (see GCBO)
% eventdata  reserved - to be defined in a future version of MATLAB
% handles    structure with handles and user data (see GUIDATA)

model = evalin ('base', 'model');
model= changeObjective(model, 'biomass_reaction');
set (handles.text2, 'String','Biomass assigned');
assignin ('base', 'model',  model);

% --- Executes on button press in fba.
function fba_Callback(hObject, eventdata, handles)
% hObject    handle to fba (see GCBO)
% eventdata  reserved - to be defined in a future version of MATLAB
% handles    structure with handles and user data (see GUIDATA)
data = evalin ('base', 'data');
model = evalin ('base', 'model');
  for j = 1:size(data,2)
        for i = 1:length(model.rxns)
            model= changeRxnBounds(model,model.rxns(i),data(i,j),'u');
            if model.rev(i)~=0
                model= changeRxnBounds(model,model.rxns(i),(-
1)*data(i,j),'l');
            else
                model= changeRxnBounds(model,model.rxns(i),0,'l');
            end
        end
        FBAsolution = optimizeCbModel(model, 'max');
            if FBAsolution.f ~= 0
                Results(:,j) = FBAsolution.x;
            end
  end
    if isempty(FBAsolution.x)
        fprintf('infeasible\n')
    else
       fileID = fopen(strcat('Results.txt'),'w');
        for ii = 1:size(Results,1)
        fprintf(fileID,'%g\t',Results(ii,:));
        fprintf(fileID,'\n');
        end
    fclose(fileID);
    end
%set (handles.text2, 'String',);
% assignin ('base', 'FBAsolution',  FBAsolution);
set (handles.text2, 'String','FBA done');
% --- Executes on button press in gpr.
function gpr_Callback(hObject, eventdata, handles)
% hObject    handle to gpr (see GCBO)
% eventdata  reserved - to be defined in a future version of MATLAB
```

```
% handles    structure with handles and user data (see GUIDATA)
Dat = uiimport()
vars = fieldnames(Dat);
for i = 1:length(vars)
   assignin('base', vars{i}, Dat.(vars{i}));
end
set (handles.text2, 'String','GPR loaded');

% --- Executes on button press in fva.
function fva_Callback(hObject, eventdata, handles)
% hObject    handle to fva (see GCBO)
% eventdata  reserved - to be defined in a future version of MATLAB
% handles    structure with handles and user data (see GUIDATA)
data = evalin ('base', 'data');
model = evalin ('base', 'model');

for j = 1:size(data,2)

        for i = 1:length(model.rxns)
            model= changeRxnBounds(model,model.rxns(i),data(i,j),'u');
            if model.rev(i)~=0
                model= changeRxnBounds(model,model.rxns(i),(-
1)*data(i,j),'l');
            else
                model= changeRxnBounds(model,model.rxns(i),0,'l');
            end
        end
            [minFlux,maxFlux] = fluxVariability(model);

                minFluxResults(:,j) = minFlux;
                maxFluxResults(:,j)= maxFlux;
end

    fileID = fopen(strcat('minFluxResults.txt'),'w');
        for ii = 1:size(minFluxResults,1)
        fprintf(fileID,'%g\t',minFluxResults(ii,:));
        fprintf(fileID,'\n');
        end
    fclose(fileID);
    fileID = fopen(strcat('maxFluxResults.txt'),'w');
        for ii = 1:size(maxFluxResults,1)
        fprintf(fileID,'%g\t',maxFluxResults(ii,:));
        fprintf(fileID,'\n');
        end
    fclose(fileID);
set (handles.text2, 'String','FVA done');

% --- Executes on button press in ko.
function ko_Callback(hObject, eventdata, handles)
% hObject    handle to ko (see GCBO)
% eventdata  reserved - to be defined in a future version of MATLAB
% handles    structure with handles and user data (see GUIDATA)
data = evalin ('base', 'data');
model = evalin ('base', 'model');

for j = 1:size(data,2)
        for i = 1:length(model.rxns)
            model= changeRxnBounds(model,model.rxns(i),data(i,j),'u');
             if model.rev(i)~=0
                model= changeRxnBounds(model,model.rxns(i),(-
1)*data(i,j),'l');
            else
                model= changeRxnBounds(model,model.rxns(i),0,'l');
            end
        end
        [grRateKO, grRateWt,grRatio, hasEffect] = singleRxnDele-
tion(model,'FBA');
        KOs(:,j) = grRateKO;
```

```
end

    fileID = fopen(strcat('KOs.txt'),'w');
        for ii = 1:size(KOs,1)
        fprintf(fileID,'%g\t',KOs(ii,:));
        fprintf(fileID,'\n');
        end
    fclose(fileID);
set (handles.text2, 'String','KOs done');

% --- Executes on button press in exit.
function exit_Callback(hObject, eventdata, handles)
% hObject    handle to exit (see GCBO)
% eventdata  reserved - to be defined in a future version of MATLAB
% handles    structure with handles and user data (see GUIDATA)
opc= questdlg('¿Desea salir del programa?','SALIR','SI','NO', 'mensajes');
if strcmp (opc,'NO')
    return;
end
clear,clc,close all
```

ANEXO 3: PUBLICACIONES

Artículos que forman parte de esta tesis doctoral

Functional proteomics outlines the complexity of breast cancer subtypes

Gámez-Pozo A, Trilla-Fuertes L, Berges-Soria J, Selevsek N, López-Vacas R, Díaz-Almirón M, Nanni P, Arevalillo JM, Navarro H, Grossmann J, Gayá Moreno F, Gómez Rioja R, Prado-Vázquez G, Zapater-Moros A, Maín P, Feliú J, Martínez del Prado P, Zamora P, Ciruelos E, Espinosa E, Fresno Vara JA

Sci Rep. 2017 Aug 30; 7(1):10100. doi: 10.1038/s41598-017-10493-w

Molecular characterization of breast cancer cell response to metabolic drugs

Trilla-Fuertes L, Gámez-Pozo A, Arevalillo JM, Díaz-Almirón M, Prado-Vázquez G, Zapater-Moros A, Navarro H, Aras-López R, Dapía I, López-Vacas R, Nanni P, Llorente-Armijo S, Arias P, Borobia AM, Maín P, Feliú J, Espinosa E, Fresno Vara JA

Oncotarget. 2018 Jan 8; 9(11):9645-9660. doi: 10.18632/oncotarget.24047. eCollection 2018 Feb 9

Computational metabolism modeling predicts risk of distant relapse-free survival in breast cancer patients

Trilla-Fuertes L, Gámez-Pozo A, Díaz-Almirón M, Prado-Vázquez G, Zapater-Moros A, López-Vacas R, Nanni P, Zamora P, Espinosa E, Fresno Vara JA

Preprint bioRxiv, doi: https://doi.org/10.1101/468595. En revisión en Future Oncology.

Probabilistic graphical models and computational metabolic models applied to the analysis of metabolomics data in breast cancer

Trilla-Fuertes L, Gámez-Pozo A, Arevalillo JM, Prado-Vázquez G, Zapater-Moros A, Díaz-Almirón M, Navarro H, Maín P, Espinosa E, Zamora P, Fresno Vara JA

Preprint bioRxiv, doi: https://doi.org/10.1101/370221. En revisión en PLOS One.

Otros artículos

Prediction of adjuvant chemotherapy response in triple negative breast cancer with discovery and targeted proteomics

Gámez-Pozo A, Trilla-Fuertes L, Prado-Vázquez G, Chiva C, López-Vacas R, Nanni P, Berges-Soria J, Grossmann J, Díaz-Almirón M, Ciruelos E, Sabidó E, Espinosa E, Fresno Vara JA

PLOS One. 2017 Jun 8; 12(6): e0178296. doi: 10.1371/journal.pone.0178296. eCollection 2017

Urothelial cancer proteomics provides both prognosis and functional information

De Velasco G, Trilla-Fuertes L, Gámez-Pozo A, Urbanowicz M, Ruiz-Ares G, Sepúlveda JM, Prado-Vázquez G, Arevalillo JM, Zapater-Moros A, Navarro H, López-Vacas R, Manneh R, Otero I, Villacampa F, Paramio JM, Fresno Vara JA, Castellano D

Sci Rep. 2017 Nov 17; 7(1):15819. doi: 10.1038/s41598-017-15920-6

Probabilistic graphical models relate immune status with response to neoadjuvant chemotherapy in breast cancer

Zapater-Moros A, Gámez-Pozo A, Prado-Vázquez G, Trilla-Fuertes L, Arevalillo JM, Díaz-Almirón M, Navarro H, Maín P, Feliú J, Zamora P, Espinosa E, Fresno Vara JA

Oncotarget. 2018 Jun 12; 9(45):27586-27594. doi: 10.18632/oncotarget.25496. eCollection 2018 Jun 12

A novel approach to triple-negative breast cancer molecular classification reveals a luminal immune-positive subgroup with good prognoses

Prado-Vázquez G, Gámez-Pozo A, Trilla-Fuertes L, Arevalillo J, Zapater-Moros A, Ferrer-Gómez M, Díaz-Almirón M, López-Vacas R, Navarro H, Maín P, Feliú J, Zamora P, Espinosa E, Fresno Vara JA

Sci Rep 2019, In press

Melanoma proteomics unravels major differences related to mutational status

Trilla-Fuertes L, Gámez-Pozo A, Prado-Vázquez G, Zapater-Moros A, Díaz-Almirón M, Fortes C, Ferrer-Gómez M, López-Vacas R, Parra Blanco V, Márquez-Rodas I, Soria A, Fresno Vara JA, Espinosa E

Preprint bioRxiv, doi: https://doi.org/10.1101/198358. En revisión en Scientific Reports.

Novel molecular classification of muscle-invasive bladder cancer opens new treatment opportunities

Trilla-Fuertes L, Gámez-Pozo A, Prado-Vázquez G, Zapater-Moros A, Díaz-Almirón M, Arevalillo JM, Ferrer-Gómez M, Navarro H, Maín P, Espinosa E, Pinto A, Fresno Vara JA

Preprint bioRxiv, doi: https://doi.org/10.1101/327114 En revisión en BMC Cancer.

Bayesian networks established functional differences between breast cancer subtypes

Trilla-Fuertes L, Zapater-Moros A, Gámez-Pozo A, Arevalillo JM, Prado-Vázquez G, Díaz-Almirón M, Ferrer-Gómez M, López-Vacas R, Navarro H, Espinosa E, Maín P, Fresno Vara JA

Preprint bioRxiv, doi: https://doi.org/10.1101/319384. En revisión en PLOS One.

SCIENTIFIC REPORTS

OPEN

Functional proteomics outlines the complexity of breast cancer molecular subtypes

Received: 4 April 2017
Accepted: 10 August 2017
Published online: 30 August 2017

Angelo Gámez-Pozo[1], Lucía Trilla-Fuertes[2], Julia Berges-Soria[1], Nathalie Selevsek[3], Rocío López-Vacas[1], Mariana Díaz-Almirón[4], Paolo Nanni[3], Jorge M. Arevalillo[5], Hilario Navarro[5], Jonas Grossmann[3], Francisco Gayá Moreno[4], Rubén Gómez Rioja[6], Guillermo Prado-Vázquez[1], Andrea Zapater-Moros[1], Paloma Main[7], Jaime Feliú[8], Purificación Martínez del Prado[9], Pilar Zamora[8], Eva Ciruelos[10], Enrique Espinosa[8] & Juan Ángel Fresno Vara[1]

Breast cancer is a heterogeneous disease comprising a variety of entities with various genetic backgrounds. Estrogen receptor-positive, human epidermal growth factor receptor 2-negative tumors typically have a favorable outcome; however, some patients eventually relapse, which suggests some heterogeneity within this category. In the present study, we used proteomics and miRNA profiling techniques to characterize a set of 102 either estrogen receptor-positive (ER+)/progesterone receptor-positive (PR+) or triple-negative formalin-fixed, paraffin-embedded breast tumors. Protein expression-based probabilistic graphical models and flux balance analyses revealed that some ER+/PR+ samples had a protein expression profile similar to that of triple-negative samples and had a clinical outcome similar to those with triple-negative disease. This probabilistic graphical model-based classification had prognostic value in patients with luminal A breast cancer. This prognostic information was independent of that provided by standard genomic tests for breast cancer, such as MammaPrint, OncoType Dx and the 8-gene Score.

Breast cancer is a major health issue in developed countries. Early diagnosis and the use of adjuvant therapies have contributed to improve survival; nevertheless, 87,000 women died of breast cancer in the European Union in 2011[1]. Knowledge of the molecular biology of breast cancer has recently challenged the way in which oncologists make decisions about systemic treatment[2].

Breast cancer is a heterogeneous disease comprising a range of entities with various genetic backgrounds. Clinical decisions are currently based on classical factors, such as the extent of the disease and the expression of hormonal receptors and human epidermal growth factor receptor 2 (HER2). Genomic classifications have also been described, the better-known encompassing four major categories: luminal A, luminal B, basal-cell and HER2-enriched[3]. Most patients included in the categories of estrogen receptor-positive (ER+)/HER2-negative (HER2-) disease with luminal A breast cancer have a favorable prognosis; however, some eventually relapse, which suggests some heterogeneity within these categories. Patients in the categories of triple-negative disease — i.e., no expression of hormonal receptors, HER2- or basal-cell disease — have a poorer prognosis[4,5].

In recent years, proteomic approaches have been incorporated into the study of clinical samples as a way to complement the information provided by classical factors and genomics. Mass spectrometry-based proteomics has emerged as preferred component of a strategy for discovering diagnostic and prognostic protein biomarkers

[1]Molecular Oncology & Pathology Lab, Institute of Medical and Molecular Genetics-INGEMM, La Paz University Hospital-IdiPAZ, Madrid, Spain. [2]Biomedica Molecular Medicine SL, Madrid, Spain. [3]Functional Genomics Center Zürich, University of Zürich/ETH Zürich, Zürich, Switzerland. [4]Department of Statistics, Biostatistics Unit, La Paz University Hospital - IdiPAZ, Madrid, Spain. [5]Operational Research and Numerical Analysis, National Distance Education University (UNED), Madrid, Spain. [6]Medical Laboratory Service, La Paz University Hospital Health Research Institute-IdiPAZ, Madrid, Spain. [7]Department of Statistics and Operations Research, Faculty of Mathematics, Complutense University of Madrid, Madrid, Spain. [8]Medical Oncology Service, La Paz University Hospital-IdiPAZ, Madrid, Spain. [9]Medical Oncology Service, Basurto Hospital, Bilbao, Spain. [10]Medical Oncology Service, Hospital 12 de Octubre (i+12) Health Research Institute, Madrid, Spain. Correspondence and requests for materials should be addressed to J.Á.F.V. (email: juanangel.fresno@salud.madrid.org)

	All	Discovery		Verification
		TNBC	ER+	All
Number of patients	106	26	80	46
Age at diagnosis (median)	54.6 (32–83)	61.2 (37–78)	54.2 (32–83)	55 (39–70)
Age at diagnosis (mean)	55.2	58.5	54.1	53.9
Tumor Size				
T1	33 (31%)	5 (19%)	28 (35%)	19 (41%)
T2	61 (58%)	19 (73%)	42 (53%)	21 (46%)
T3	10 (9%)	2 (8%)	8 (10%)	6 (13%)
T4	1 (1%)	0 (0%)	1 (1%)	0 (0%)
Multifocal	1 (1%)	0 (0%)	1 (1%)	0 (0%)
Tumor Grade				
G1	12 (11%)	0 (0%)	12 (15%)	6 (13%)
G2	33 (31%)	4 (15%)	29 (36%)	22 (48%)
G3	41 (39%)	20 (77%)	21 (26%)	12 (26%)
Unknown	20 (19%)	2 (8%)	18 (23%)	6 (13%)
Lymph node status				
N0	0 (0%)	0 (0%)	0 (0%)	0 (0%)
N1	71 (67%)	17 (65%)	54 (68%)	39 (85%)
N2	35 (33%)	9 (35%)	26 (32%)	7 (15%)
Chemotherapy				
No anthracyclines	34 (32%)	11 (42%)	23 (29%)	0 (0%)
Anthracyclines	63 (59%)	12 (46%)	51 (64%)	66 (100%)
Anthracyclines + taxanes	9 (9%)	3 (12%)	6 (7%)	0 (0%)

Table 1. Patient's characteristics.

as well as for establishing new therapeutic targets[6]. Although these investigations are encouraging[7,8], the number of tumor biomarkers discovered with this approach is still limited[9]. MicroRNAs are key regulators in the genesis and progression of cancer. MicroRNA profiling, together with genomics and proteomics, could lead to unraveling regulatory networks of biological processes related to cancer[10].

In this study, we used high-throughput proteomics and microRNA profiling to characterize two subtypes of breast cancer with various prognoses: ER+/progesterone receptor-positive (PR+) HER2- breast cancer and triple-negative breast cancer (TNBC). We applied probabilistic graphical models and flux balance analyses to explore molecular differences between these subtypes to unveil differences not detected by immunohistochemistry or genomics.

Results

Patient characteristics. A total of 106 patients with breast cancer from two different hospitals were included in the discovery cohort. All the patients had node-positive disease, all the tumors were negative for HER2 and all had received adjuvant chemotherapy and hormonal therapy for patients with ER+ disease (patients showing estrogen and/or progesterone receptor expression). Forty-six additional patients from a third hospital with ER+ disease and nodal involvement were eligible for the verification cohort: all had received anthracycline-based adjuvant chemotherapy followed by hormone therapy (Table 1 and Sup. Fig. S1).

Protein extraction and shotgun-mass spectrometry analyses of formalin-fixed, paraffin-embedded breast cancer tumors. After mass spectrometry (MS) workflow, 25 TNBCs and 71 ER+ tumors from the discovery cohort were analyzed. Raw data normalization was performed as previously described[10]. Four samples were excluded due to poor protein extraction and six were excluded due to data quality. Of 3,239 protein groups identified using Andromeda, 1095 presented at least two unique peptides and detectable expression in at least 75% of the samples in either the ER+ or TNBC groups. No decoy protein passed through these additional filters. Label-free quantification data were obtained using MaxQuant as previously described[10].

Protein expression analyses of breast cancer tumors. Protein expression values were analyzed using Significance Analysis of Microarrays (SAM). A total of 224 proteins were differentially expressed between the ER+ and TNBC samples with a false discovery rate (FDR) <5% (Sup. Table S1). Hierarchical clustering analysis split the samples into two main clusters: cluster I comprised 70.4% of ER+ tumors (labeled ER-true), and cluster II included both ER+ and triple-negative (TN) tumors. The ER+ tumors included in cluster II, representing 29.6% of all ER+ tumors, were labeled as TN-like tumors. The distant metastasis-free survival (DMFS) rate at 5 years was 88.2% for ER-true and 71.4% for TN-like patients ($p = 0.21$). The clinical evolution of TN-like breast cancer was similar to that of TNBC (DMFS rate at 5 years 65.4%, $p = 0.7$) (Fig. 1).

Characterization of ER-true and TN-like subtypes. A significance analysis of microarrays (SAM), excluding TNBC tumors, was performed to further characterize ER-true and TN-like subtypes. We found 44 proteins

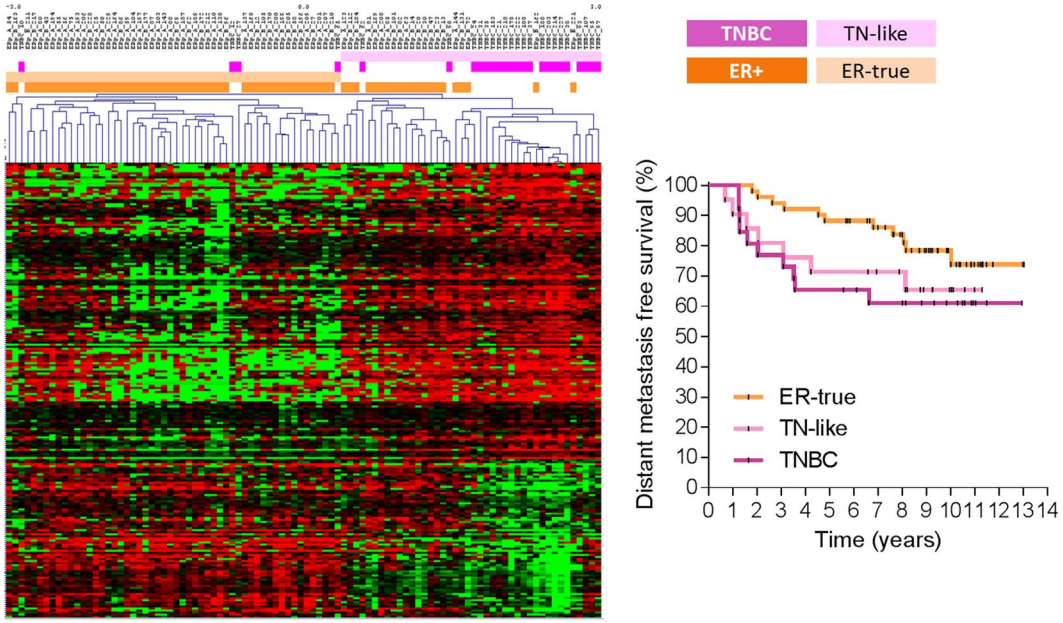


Figure 1. ER-true/TN-like subtype definition and characterization. Left panel: Hierarchical clustering analysis from 224 proteins identified by SAM analysis between ER+ and TNBC tumors with FDR < 5%. Right panel: Kaplan-Meier analysis showing survival for ER-true, TN-like and TNBC tumors (n = 51, 21 and 26, respectively; p = 0.17).

showing differential expression between both subgroups, with an FDR < 5% (Sup. Table S2 and Sup. Fig. 2). Four proteins presented deleted records in Uniprot and were excluded. Among the proteins with higher expression in ER-true tumors, we found 7 extracellular small leucine-rich canonical proteoglycans (SLRPs) (biglycan, decorin, asporin, lumican, prolargin, fibromodulin and osteoglycin), three proteins produced by mast cells (cathepsin G, mast cell carboxypeptidase A and chymase), COEA1, PRDBP, and both the PIP and ZA2G proteins. On the other hand, TN-like tumors showed greater expression of HS90B and STIP1 from the chaperone pathway, EF2 and THEM6 proteins. Gene ontology analyses showed that proteins defining the TN-like subtype were related to cell adhesion processes (Sup. Table S3). Regarding clinical factors, we found that TN-like tumors showed higher molecular grade (G1-2 vs. G3, p = 0.03). No differences between ER-true and TN-like tumors regarding age at diagnosis, tumor size, number of affected nodes, and ER, PR or Ki67 pathological assessment were found.

MicroRNA expression analysis of breast cancer tumors. MicroRNA expression profiling was available for 42 ER-true and 23 TN-like tumors from the discovery cohort. One microRNA was excluded from subsequent analyses due to absence of expression in most of the samples. Nine microRNAs showed significant higher expression in the ER-true compared with the TN-like tumors (p < 0.05; FDR < 5%) (Sup. Fig. S3).

Systems biology of ER+ breast cancer. Both label-free protein quantification and microRNA expression data were available for 16 TNBC and 63 ER+ tumors from the discovery cohort. A probabilistic graphical model was constructed with these values as previously described[10]. Differences in functional node activity between ER-true and TN-like tumors were found (Figs 2 and S4). These differences were corroborated in the external dataset (p < 0.05), except for the protein synthesis node. All metabolism and mitochondria nodes present higher activity in TN-like tumors. The "metabolism A" node includes proteins related to glutamine and glucose metabolism and LDHB. The "metabolism B" node includes GAPDH, PGK1, LDHA and pyruvate kinase proteins, among others; and also miR-449a, whose expression showed a negative correlation with the functional node activity (Sup. Fig. S6). The "mitochondria A" node includes proteins related to the mitochondrial oxidation/reduction process, whereas the "mitochondria B" node comprises tricarboxylic acid (TCA) cycle proteins. The "ECM & focal adhesion" node showed higher activity in ER-true tumors, and includes miR-139-5p, miR-149, miR-766, miR-342, miR-214* and miR-31. Both miR-214* and miR-31 expression showed positive correlation with functional node activity (Sup. Fig. S5). The "response and membrane" node includes proteins related to cellular response to external stimuli and cholesterol homeostasis, and shows higher activity in ER-true tumors. The "proteasome" node includes proteins from the proteasome core complex, and showed higher activity in TN-like tumors. This functional node includes miR-489 and miR-99a, although no correlation was found between their expression and functional node activity.

Flux balance analysis of breast tumors. We performed a flux balance analysis (FBA) using the E-Flux algorithm to evaluate the impact of the proteomics profile on tumor growth capability[11]. Our Recon 2-based model includes 7440 reactions, from which we found gene-protein-reaction (GPR) rule values mediating 1085 reactions. All the tumors fulfilled the Warburg effect, redirecting pyruvate generated by glycolysis and

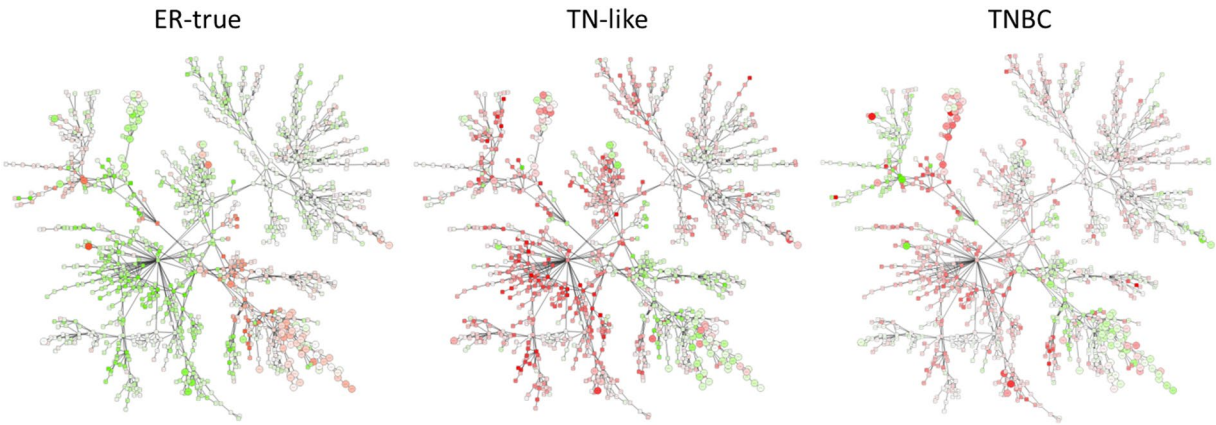


Figure 2. Protein- and miRNA-based probabilistic graphical model. Probabilistic graphical model showing protein (squares) and miRNA (circles) mean expression in each sample type. Color range from -2-fold change (green) to 2-fold change (red). White means no change between groups. ER-true subtype is compared with TN-like subtype and *vice versa*. TNBC type is compared with all ER+ tumors.

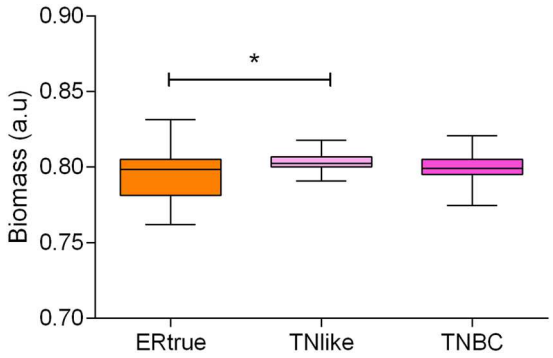


Figure 3. Tumor growth rate predicted by flux balance analysis. FBA results for ER-true, TN-like and TNBC tumors (n = 51, 21 and 26, respectively; *$p < 0.05$).

glutaminolysis to lactic fermentation through lactate dehydrogenase. The predicted tumor growth rate was higher in both the TN-like and TNBC tumors compared with the ER-true tumors (Fig. 3).

Targeted proteomics of TN-like/ER-true subtypes. To corroborate the prognostic value of the TN-like/ER-true classification, 33 proteins differentially expressed between TN-like and ER-true subtypes were assessed, using a targeted proteomics approach via selected reaction monitoring (SRM) in a new cohort comprising 46 ER+ breast cancer tumors (Table 1)[12]. One sample was excluded due to poor protein extraction and two due to data quality. Nineteen samples from the discovery cohort were also tested. SRM was able to detect differences between ER-true and TN-like samples from the discovery cohort (Sup. Fig. S6). An ER-true/TN-like classifier, including 14 proteins, was used to assign new samples to ER-true or TN-like (sup. info). DMFS rates at 5 years were 81.6% and 57.8% for the ER-true and TN-like groups, respectively ($p < 0.17$) (Fig. 4).

Assessing ER-true/TN-like subtypes using a meta-genomics external dataset. We used gene expression data from 1296 breast cancer tumors, obtained from public repositories, as an independent cohort to validate the prognostic value of the ER-true/TN-like stratification[13,14]. Among them, 935 tumors were ER+ and had follow-up information available. Tumors were labeled as ER-true or TN-like using 35 of 44 proteins from SAM analyses. Survival analyses using 421 tumors with ER+ and node positive characteristics showed that DMFS rates were 81.8% and 72.5% for the ER-true and TN-like groups, respectively ($p < 0.005$, HR = 0.5769, Sup. Fig. S7).

ER-true/TN-like subtypes and breast cancer molecular subtypes. We applied our TN-like classifier to the entire population and performed survival analyses independently for each breast cancer molecular subtype[3]. ER-true/TN-like subtyping provided additional prognostic information in luminal A tumors, but not in luminal B, basal or HER2-enriched tumor subtypes (Fig. 5).

ER-true/TN-like subtypes and molecular prognostic signatures. The clinical utility of the ER-true/TN-like subtypes was evaluated in combination with three prognostic gene signatures: the 70-gene signature[15],

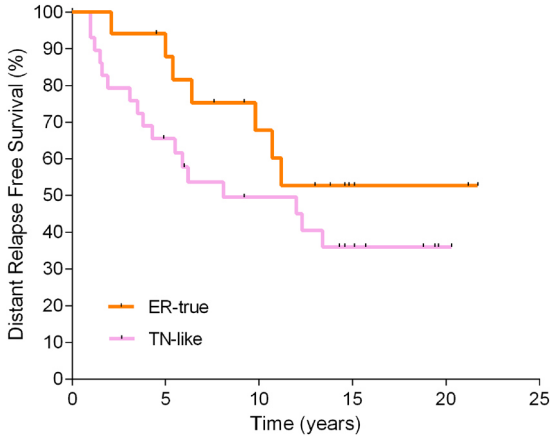


Figure 4. SRM validation of new subtypes. Kaplan-Meier analysis showing survival rates for ER-true and TN-like tumors on the basis of SRM data (n = 17 and 29, respectively).

the Recurrence Score[16] and the 8-gene Score[17]. The prognostic value of the three tests in 935 patients with ER+ tumors was corroborated, followed by the application of the ER-true/TN-like class predictor (Fig. 6). The TN-like tumors were associated with a lower DMFS compared with ER-true tumors in each low-risk category, defined by prognostic signatures (Sup. Table S4). The high-risk categories were not further refined through ER-true/TN-like subtyping. The multivariate analyses, including each prognostic signature and the ER-true and TN-like subtypes, showed that the ER-true and TN-like subtypes were related to prognosis, independent of the prognostic gene signatures (Table 2). Multivariate analyses including the ER-true/TN-like subtypes and available clinical variables (grade and N) showed that both grade and ER-true/TN-like subtypes, along with lymph node status, provided significant and independent prognostic information.

Discussion

In this study, a new subtype of breast cancer was identified using a proteomics approach. The clinical classification of breast cancer does not fully reflect cancer heterogeneity; thus, individuals receiving the same diagnosis can have markedly different outcomes. Genomics and proteomics approaches complement the information provided by routine determinations, and coupled with new data analysis techniques, they help to expand the information obtained. In this case, information provided by pure protein expression was organized into functional nodes involving specific biological processes and pathways. The new TN-like ER+ subtype defined has molecular features common with TNBC tumors and exhibits a similar clinical evolution. Patients with either TN-like or TNBC tumors have shorter DMFS than patients with ER-true breast cancer. Both SRM verification and meta-validation confirmed the findings obtained in the discovery series. These results might help to explain why the prognosis of patients with ER+ breast cancer is not uniformly favorable.

ER-true tumors present molecular features that could explain the favorable prognosis of this subtype, such as increased expression of proteins related to cell adhesion and greater activity of the "ECM & focal adhesion" node. Increased expression of decorin and lumican in breast cancer is associated with lower tumor size, decreased risk and rate of relapse, positive ER/PR status and better survival[18,19]. A stromal gene set including DCN and FBLN1 genes has demonstrated prognostic value independent of clinical information and a proliferation gene set[20]. COEA1, asporin, osteoglycin and lumican showed increased expression in low-risk vs. high-risk tumors defined by MammaPrint[21]. With regard to miRNAs included in the "ECM & focal adhesion" node, miR-342 expression correlates with ER expression and tamoxifen sensitivity in breast tumors[22–24]. Both miR-149 and miR-342 have been included in a prognostic signature for breast cancer[25]. Our results suggest that miR-31 and miR-214* could be indirect regulators of cell attachment function in breast tumors. These results indicate that ER-true tumors harbor a limited metastatic potential compared with TN-like tumors.

There is more limited information on some other molecular features defining ER-true tumors. This subtype has high expression for proteins produced by mast cells, related to ER and PR positivity, low-grade and a good prognosis in breast cancer[26–29]. High expression levels of PIP and AZGP1 genes have been related to a good prognosis and correlate with ER, PR and AR expression[21,30–38]. ZA2G is part of a panel of 13 proteins predicting recurrence in breast cancer, showing decreased expressions when recurrence occurred[39]. PRDBP protein appears to dictate the balance between ERK and Akt signaling with consequences for cell metabolism (induction of Warburg metabolism), apoptosis and cell proliferation[40]. Loss of the 11p15 region, where the PRKCDBP gene is located, is common in breast cancer metastases[41].

TN-like tumors showed molecular features associated with an unfavorable prognosis. High HSP90AB1 expression is related to poor overall survival and with an increased distant metastasis relapse rate in breast cancer[42,43] which is consistent with our results, showing a higher expression of this protein in TN-like tumors. HS90B has been included in a panel of 13 proteins predicting recurrence in breast cancer[39]. STIP1 interacts with HS90B in the folding of a number of proteins, including the androgen and estrogen receptors[44,45]. Additionally, greater expression of eEF2 was significantly associated with node positivity in breast cancer[46].

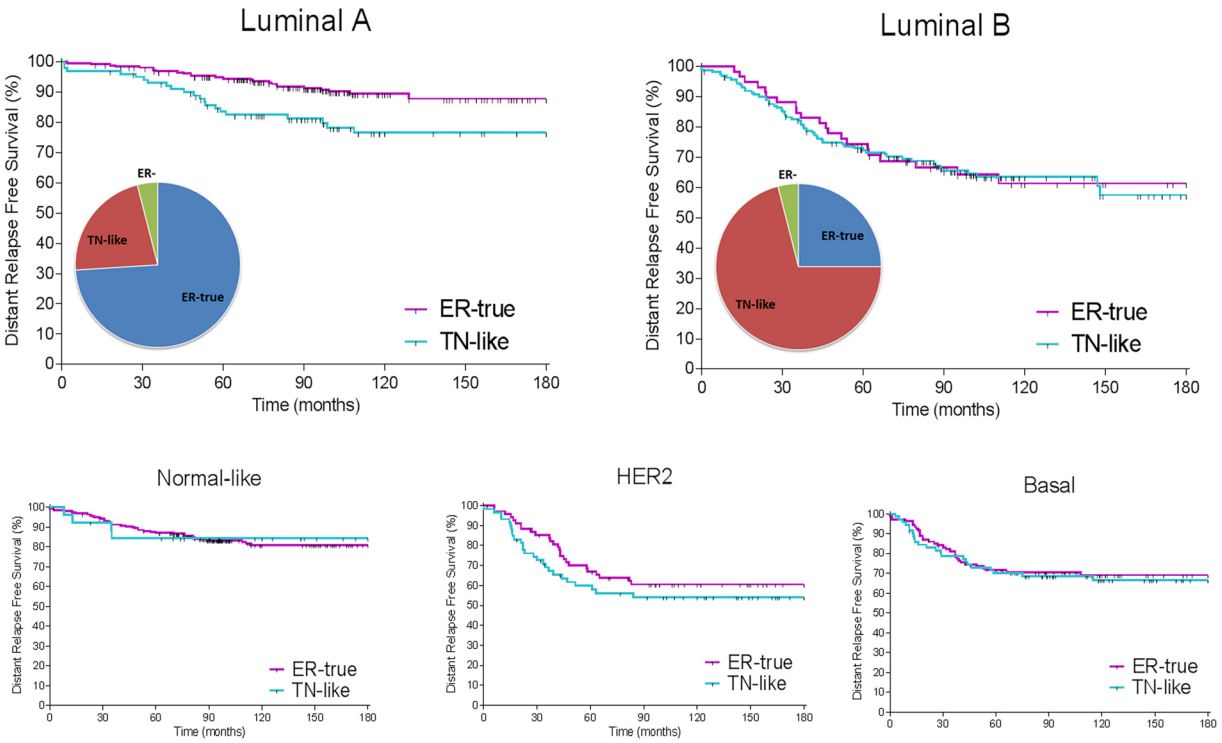


Figure 5. Prognostic value of ER-true/TN-like subtype within breast cancer molecular subtypes. Kaplan-Meier analysis showing ER-true and TN-like tumor survival rates in luminal (**A**) (left panel: ER-true n = 262, TN-like n = 101) and luminal (**B**) (right panel: ER-true n = 59, TN-like n = 164) subtypes.

On the other hand, all metabolism and mitochondria nodes had higher activity in the TN-like subtype. The "mitochondria B" and "metabolism B" nodes include proteins related to the TCA cycle and glycolysis, respectively, suggesting that both TN-like and TNBC tumors are highly glycolytic. The TN-like tumors showed high activity in both the "metabolism A" and "mitochondria A" nodes compared with the ER-true and TNBC tumors, suggesting a unique metabolic profile for the TN-like subtype. FBA indicates that all breast cancer types fulfill the Warburg hypothesis and that glutamine-derived αKG refuels the TCA cycle (anaplerosis) and maintains constant levels of biosynthetic precursors, while the surplus turns to lactate[47]. However, ER-true tumors had a predicted growth rate significantly lower than TN-like and TNBC tumors, both of which had comparable growth rates.

Molecular differences between TN-like and ER-true tumors resemble those previously described between ER+ and TNBC tumors[10]. SAM analysis identified 44 proteins differentially expressed between both subtypes, 24 of which were also differentially expressed between ER+ and TNBC samples. Moreover, miR-139-5p, miR-149, miR-449a and miR-342 were overexpressed in ER+ tumors with regard to TNBCs[10, 24, 48, 49]. Interestingly, we found equivalent differences in the "ECM & focal adhesion," "metabolism B," "mitochondria B" and "protein synthesis" nodes when comparing ER-true versus TN-like tumors and ER+ versus TNBC tumors. Differences in the "protein synthesis" node could not be confirmed in the external dataset in both analyses, suggesting that some features observed at the protein level do not appear at the gene expression level[10]. On the other hand, no differences regarding the "proliferation" node activity between ER-true and TN-like samples were found, although they were present between the ER+ and TNBC tumors. We also found differences not described between ER+ and TNBC tumors: the "mRNA processing" and "protein transport" nodes showed higher activity in ER-true tumors, whereas the "response and membrane" node had higher activity in TN-like tumors.

The TN-like subtype added prognostic information in luminal A disease but not in the other molecular subtypes. Likewise, the TN-like subtype further subdivided low-risk categories defined by gene signatures, such as the 70-gene Score, Recurrence Score and 8-gene Score. These gene signatures are related to cell proliferation, whereas the TN-like subtype primarily depends on other drivers, such as cell attachment and metabolism, thus providing complementary information[50]. New molecular information could improve the accuracy of gene signatures and help to determine the best treatment for patients with ER+ breast cancer. Additionally, the TN-like subtype prognostic information is independent of that provided by clinical variables such as lymph node status and grade. Adjuvant treatment of breast cancer is determined by two main factors: risk of relapse and the molecular characteristics of the tumor. Molecular tools developed in this setting — such as MammaPrint, OncoType or the 8-gene Score — have attempted to optimize the use of adjuvant chemotherapy, which is toxic and benefits a limited number of patients. Patients in the low-risk categories of these gene tests do not require chemotherapy, but our results indicate that these low-risk categories can be further subdivided. The presence of a TN-like subtype worsens the outcome; therefore, chemotherapy should be considered in these patients. The recommendation

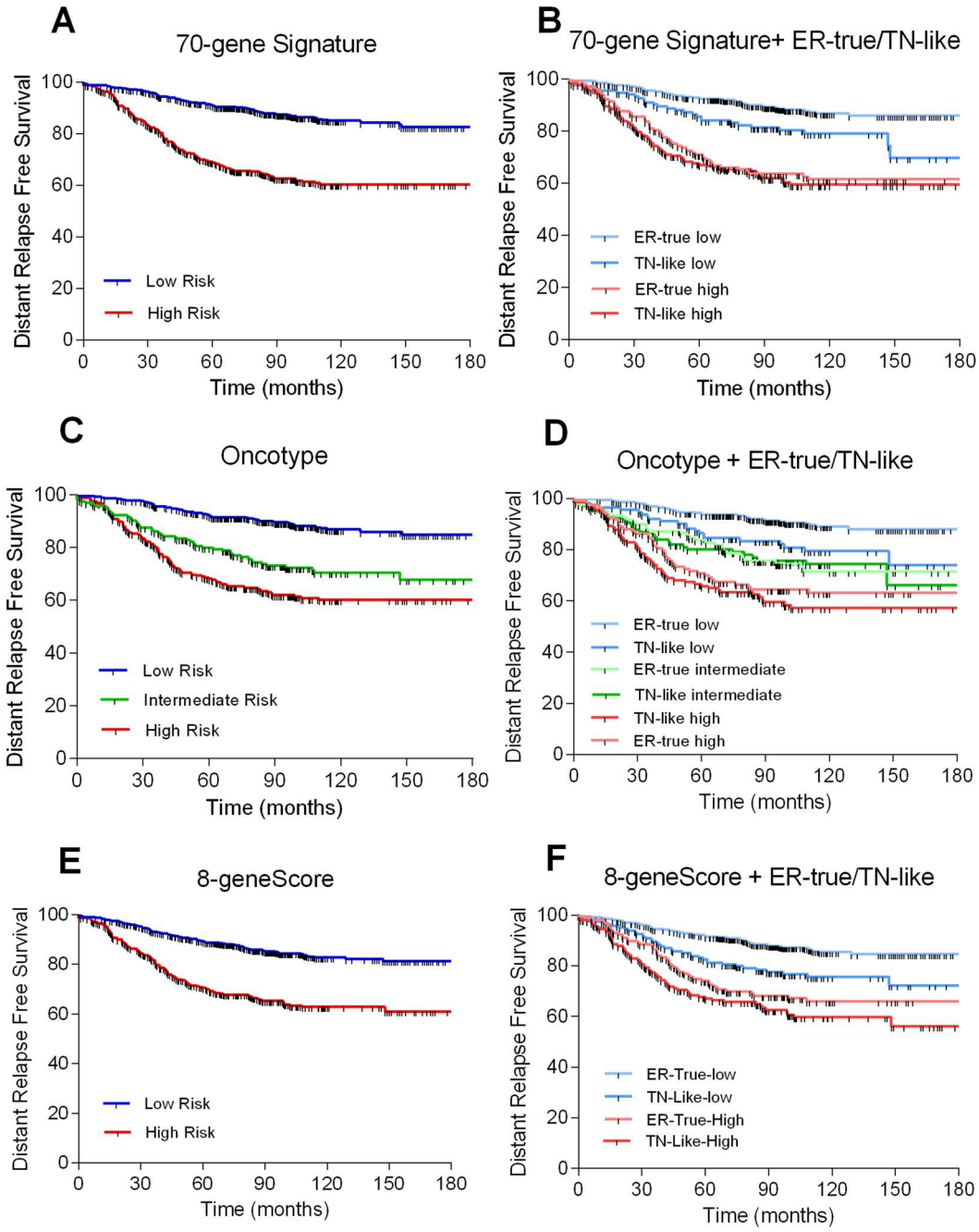


Figure 6. ER-true/TN-like subtype and prognostic signatures. Kaplan-Meier analysis showing survival rates of risk groups defined by prognostic gene signatures and ER-true/TN-like subtypes. (**A**) 70-gene Signature: Low risk = 586; High risk = 349; $p < 0.0001$; HR = 3.24 (2.73–4.85). (**B**) 70-gene Signature and ER-true/TN-like subtypes: Low risk/ER-true = 449; High risk/ER-true = 154; Low risk/TN-like = 137; High risk/TN-like = 195; $p < 0.0001$. (**C**) Recurrence Score: Low risk = 472; Intermediate risk = 195; High risk = 268; $p < 0.0001$. (**D**) Recurrence Score and ER-true/TN-like subtypes: Low risk/ER-true = 358; Intermediate risk/ER-true = 120; High risk/ER-true = 268; Low risk/TN-like = 125; Intermediate risk/TN-like = 108; High risk/TN-like = 143; $p < 0.0001$. (**E**) 8-gene Score: Low risk = 610; High risk = 325; $p < 0.0001$; HR = 2.61 (2.19–3.94). (**F**) 8-gene Score and ER-true/TN-like subtypes: Low risk/ER-true = 445; High risk/ER-true = 158; Low risk/TN-like = 165; High risk/TN-like = 167; $p < 0.0001$.

would be valid for luminal A tumors having features of the TN-like subtype, which could contribute to reducing the number of relapses in this population.

The TN-like subtype provided prognostic information in ER+ disease not only with the original proteomics approach, but also with other techniques, including the translation of proteins back to gene expression. This result supports the robustness of this new breast cancer subtype. In addition, some of the components defining the subtype could become potential therapeutic targets in the future. Hormonal receptors and HER2 are the only

Univariate analysis		
	p-value	HR
ER-true/TN-like subtype	$<10^{-4}$	1.911
70-gene Signature	$<10^{-4}$	3.239
Recurrence Score	$<10^{-4}$	1.929
8-gene Score	$<10^{-4}$	2.605
Multivariate analysis TN-like subtype and clinical variables		
ER-true/TN-like subtype	0.022	1.374
Grade (1 + 2 vs. 3)	$>10^{-4}$	1.555
N	0.005	1.481
Multivariate analysis TN-like subtype and prognostic signatures		
ER-true/TN-like subtype	0.05	1.329
70-gene Signature	$>10^{-4}$	2.948
ER-true/TN-like subtype	0.011	1.441
Recurrence Score	$>10^{-4}$	1.829
ER-true/TN-like subtype	0.002	1.544
8-gene Score	$>10^{-4}$	2.336

Table 2. Univariate and multivariate analyses including clinical variables, prognostic signatures and the TN-like subtype.

molecular features allowing targeted therapy in breast cancer. Gene subtyping into luminal A, luminal B, basal and HER-2 enriched groups has not revealed other features that can be used to develop new drugs. A proteomics approach unravels molecular processes not detected by genomics, with the advantage that proteins are the real effectors of genomic changes.

Our study has some limitations. The discovery series was limited to patients with node-positive disease, who have a poorer outcome than their node-negative counterparts. However, the meta-validation series is more heterogeneous regarding clinical stage, which suggests that the TN-like subtype is a clinical entity and not just a marker of advanced disease. Also, relevant clinical differences in the discovery and verification cohorts did not reach the statistical boundary due to the limited sample size and the fact that many relapses in this group appeared after 5 years of follow-up. This problem was overcome in the *in-silico* series, which is more representative of a population of patients with breast cancer. On the technical side, despite the informative value of proteomics, there is still room for improvement in the number of proteins detected. Moreover, SRM assays are complex to develop and analyze in comparison with other platforms such as quantitative polymerase chain reaction (qPCR), and its use in the clinical routine is still challenging. Finally, these results should be validated in additional cohorts to evaluate the TN-like subtype robustness.

High-throughput proteomics generate clinically useful protein-based molecular profiles, which can complement information provided by gene expression analysis. In this study, a proteomics approach allowed the identification of a new subtype of breast cancer using FFPE samples. The molecular characteristics of this new subgroup have been assessed using probabilistic graphical models. This subtype is included in the group of hormonal receptor-positive, HER2-negative tumors, but has molecular features and a poor clinical outcome similar to that of TNBC. This new TN-like subtype has the capability to add prognostic information to current clinical practice. Because proteins are the final effectors of genes, some proteins and biological processes defining TN-like tumors could become therapeutic targets. This possibility should be further explored in future studies.

Methods

Sample selection. A total of 106 patients with breast cancer were included in the discovery cohort. FFPE samples were retrieved from the I+12 Biobank (RD09/0076/00118) and from the IdiPAZ Biobank (RD09/0076/00073), both integrated in the Spanish Hospital Biobank Network (RetBioH; www.redbiobancos.es). Forty-six patients were included in the verification cohort, and FFPE samples were retrieved from the Basque Biobank/O+EHUN (RD09/0076/00140). Informed consent was obtained from all the patients. All the experiments were performed in accordance with relevant guidelines and regulations. The histopathological features of each sample were reviewed by an experienced pathologist to confirm diagnosis and tumor content. Eligible samples included at least 50% tumor cells. Approvals from the Ethics Committees of Hospital Doce de Octubre, La Paz University Hospital and Euskadi were obtained for the conduct of the study.

Total protein preparation and digestion. Proteins were extracted from FFPE samples as previously described[51]. Briefly, FFPE sections were deparaffinized in xylene and washed twice with absolute ethanol. Protein extracts from the FFPE samples were prepared in 2% sodium dodecyl sulfate (SDS) buffer using a protocol based on heat-induced antigen retrieval[52]. Protein concentration was determined using the MicroBCA Protein Assay Kit (Pierce-Thermo Scientific). Protein extracts (10 µg) were digested with trypsin (1:50) and SDS was removed from digested lysates using Detergent Removal Spin Columns (Pierce). Peptide samples were further desalted using ZipTips (Millipore), dried, and resolubilized in 15 µL of a 0.1% formic acid and 3% acetonitrile solution before MS analysis.

Liquid chromatography - mass spectrometry shotgun analysis. The samples were analyzed on an LTQ-Orbitrap Velos hybrid mass spectrometer (Thermo Fischer Scientific, Bremen, Germany) coupled to a NanoLC-Ultra system (Eksigent Technologies, Dublin, CA, USA) as previously described previously[10]. Briefly, after separation, peptides were eluted with a gradient of 5% to 30% acetonitrile in 95 minutes. The mass spectrometer was operated in data-dependent mode (DDA), acquiring a full-scan MS spectra (300–1700 m/z) at a resolution of 30,000 at 400 m/z after accumulation to a target value of 1,000,000, followed by collision-induced dissociation (CID) fragmentation on the 20 most intense signals per cycle. The samples were acquired using internal lock mass calibration on m/z 429.088735 and 445.120025. The acquired raw MS data were processed by MaxQuant (version 1.2.7.4)[53], followed by protein identification using the integrated Andromeda search engine[54]. Briefly, spectra were searched against a forward UniProtKB/Swiss-Prot database for human, concatenated to a reversed decoyed FASTA database (NCBI taxonomy ID 9606, release date 2011-12-13). The maximum FDR was set to 0.01 for peptides and 0.05 for proteins. Label-free quantification was calculated on the basis of the normalized intensities (LFQ intensity). Quantifiable proteins were defined as those detected in at least 75% of samples in at least one type of sample (either ER+ or TNBC samples) showing two or more unique peptides. Only quantifiable proteins were considered for subsequent analyses. Protein expression data were log2 transformed, and missing values were replaced using data imputation for label-free data, as explained in Deeb *et al.*[55], using default values. Finally, protein expression values were z-score transformed. Batch effects were estimated and corrected using ComBat[56]. All the mass spectrometry raw data files acquired in this study can be downloaded from Chorus (http://chorusproject.org) under the project name *Breast Cancer Proteomics*.

RNA extraction and MicroRNA expression. RNA isolation from the FFPE tumor specimens and microRNA expression profiling was performed as previously described[10]. Briefly, microRNA expression profiling was obtained using a custom TaqMan Array MicroRNA Card (Applied Biosystems) containing 95 FFPE-reliable assays, including four housekeeping miRNAs identified used NorMean[57]. ΔCq values were normalized using two reference miRNAs (hsa-let-7d and hsa-let-7g).

Differential expression analysis of label-free proteomics and microRNA profiling. SAM[58] was performed to find differentially expressed proteins and miRNAs between sample groups with an FDR below 5%. Hierarchical clusters were constructed with the differentially expressed proteins or miRNAs between predefined samples groups identified by SAM, using Pearson's correlation and the average-linkage method.

Functional network construction. A functional network to associate miRNAs and protein expression profiles was constructed as previously described[10]. Briefly, we chose probabilistic graphical models compatible with high-dimensionality. The result is an undirected graphical model with a local minimum Bayesian Information Criterion (BIC)[59]. Methods are implemented in the open-source statistical programming language R[60]; in particular, the functions *minForest* and *stepw* in the *gRapHD* package[61]. To identify functional nodes within the network, we split it into several branches or functional nodes. We then used gene ontology analyses to investigate which function or functions were overrepresented in each branch. To measure the functional activity of each functional node, we calculated the mean expression of all the proteins included in one branch related to a specific function. Differences in functional node activity were assessed by class comparison analyses.

Gene ontology analyses. Protein-to-gene ID conversion was performed using Uniprot (http://www.uniprot.org) and DAVID[62,63]. Gene ontology analyses were performed using the functional annotation chart tool provided by DAVID. We used "*homo sapiens*" as a background list and selected only GOTERM-FAT gene ontology categories and Biocarta, KEGG and Panther pathways.

Flux balance analyses. Flux balance analysis (FBA) is a widely used approach for studying biochemical networks by calculating the flow of metabolites through the network, including 7440 reactions from Recon 2[64]. With this method, it is possible to predict the growth rate of an organism or the rate of production of a metabolite[65]. The estimation of the GPR rule values was performed using a variation of the method described by Barker *et al.*[66]. The mathematical operations used to calculate the numerical value were the sums for "OR" expressions and minimums for "AND" expressions. Finally, the GPR rule values were normalized, dividing by the maximum value in each tumor, and were included in the Recon 2 model using the E-Flux algorithm[11]. Normalized GPR rule values have been used to establish both lower and upper reaction bounds if the reaction is reversible. If the reaction is irreversible, low bound is set to 0 in all cases. To calculate biomass production, the biomass objective function included in Recon 2 was optimized. FBA was performed using the COBRA Toolbox available for MATLAB[67].

Selected reaction monitoring analyses. The SRM design was based on both experimental data from our shotgun analysis and the PeptideAtlas[68]. SRM-triggered MS2 was performed on a QTRAP 5500 instrument (ABSciex, Concord, Ontario), and SRM measurements were analyzed on a TSQ Vantage Triple Quadrupole Mass Spectrometer (ThermoFisher, San Jose, CA, USA), both equipped with a nanoelectrospray ion source. Chromatographic separations of peptides were performed on a NanoLC-2D HPLC system (Eksigent, Dublin, CA) coupled to a 15-cm fused silica emitter, 75-μm diameter, packed with a ReproSil-Pur C18-AQ 120 A and 1.9-μm resin (Dr. Maisch HPLC GmbH). Peptides were loaded on the column from a cooled (4 °C) Eksigent autosampler and separated with a linear gradient of acetonitrile/water, containing 0.1% formic acid, at a flow rate of 300 nl/min. A gradient from 5% to 35% acetonitrile in 40 minutes was used. For the SRM-triggered MS2 measurements, MS2 spectra were recorded upon detection of an SRM trace above a threshold of 1000 ion counts. An average of 100 transitions (scan time 10 ms/transition) per run was used and Q1 and Q3 were obtained at 0.7 unit mass resolution. MS2 spectra were recorded in enhanced product ion (EPI) mode for the highest MRM transitions, using dynamic fill time, Q1 resolution unit, scan speed 10,000 amu/s, m/z range 300–1000. Collision

energies used for both acquisition modes were calculated according to the formulas: CE = 0.044 * m/z + 5.5 and CE = 0.051 * m/z + 4 (CE: collision energy; m/z: mass-to-charge ratio of the precursor ion) for doubly and triply charged precursor ions, respectively. In SRM, the mass spectrometer was operated in SRM scan mode, in which Q1 and Q3 were obtained at 0.7 unit mass resolution. Collision energies for each transition were calculated according to the following equations: CE = 0.034 * (m/z) + 3.314 and CE = 0.044 * (m/z) + 3.314 for doubly and triply charged precursor ions, respectively. Three SRM transitions were monitored for each endogenous (light) and internal standard (heavy) peptide. SRM data were processed using SRM skyline software[69]. Peptides with the following criteria were used for the quantification: (i) correlation between ion ratios obtained for the heavy and the light form; (ii) correlation between the ion ratios obtained for both forms and the ion ratios obtained in the MS/MS spectra present in the SRM spectral library; and (iii) transition intensities of the heavy and the light form of >10. The three transitions for each heavy-light pair were used to quantify the peptide unless signals of coeluting interferences were detected. Punctual measurements for light peptides below the background measurement value were ignored. A light/heavy peptide ratio was calculated for all transitions. Protein expression values were calculated by the median expression from the three transitions for each heavy-light pair of their peptides.

Development of classifiers. We developed protein expression-based signatures to predict the class of future samples using the compound covariate predictor. The model incorporates proteins that were differentially expressed among classes at the 0.05 significance level as assessed by the random variance t-test[70], with protein-to-gene ID positive in the meta-validation dataset (see below). We estimated the prediction error of each model using leave-one-out cross-validation (LOOCV)[71]. For each LOOCV training set, the entire model building process was repeated, including the gene selection process. We also evaluated whether the cross-validated error rate estimate for a model is significantly less than the random prediction. The class labels were randomly permuted and the entire LOOCV process was repeated. The significance level is the proportion of the random permutations that gave a cross-validated error rate no greater than the cross-validated error rate obtained with the original data. The same workflow was performed using the SRM data. For more details, see the Simon R and Lam A. BRB-ArrayTools User Guide, version 3.2. BRB-ArrayTools v4.2.1, developed by R. Simon and A. Peng.

External dataset validation. A total of 1296 primary breast carcinoma data were collected from two independent datasets[13,14]. The Guedj dataset and associated clinical annotations were downloaded from the ArrayExpress Archive (http://www.ebi.ac.uk/arrayexpress/). The Miller dataset and associated clinical annotations were downloaded from the Cancer Research website. Batch effects were corrected using ComBat[56]. Protein-to-gene ID was performed using Uniprot (http://www.uniprot.org) and DAVID[62,63]. All the probes in the dataset for each gene were retrieved. Probes with higher coefficients of variation were selected when multiple probes were found for a single gene, then expression values of each gene were z-score transformed. Samples with clinical characteristics similar to those in our discovery cohort were then assigned to various groups using the developed predictor. The 70-gene signature[15], Recurrence Score[16] and 8-gene Score predictors were calculated for all the samples in the dataset as described previously[14,17,72]. Molecular subtype annotation was performed using the Single Sample Predictor described by Hu *et al.*[72,73]. To apply protein expression-based signatures to gene expression values, per-gene normalization was applied as previously described[17].

Statistical analyses and software suites. Survival curves were estimated using a Kaplan-Meier analysis and compared with the log-rank test, using DMFS at 5 years as the end point. Univariate and multivariate Cox proportional hazard analyses were also employed to evaluate the defined prognosis predictors. Correlations were assessed using Pearson's r and linear regression. The SPSS v16 software package, GraphPad Prism 5.0 and R v2.15.2 (with the *Design* software package 0.2.3) were used for all the statistical analyses. Correlation between node activity and microRNA expression was evaluated using linear regression analyses and Pearson's correlation. Comparisons between different populations' characteristics were assessed using Fisher's exact test, the chi-squared test or the Mann-Whitney test as appropriate. All p-values were two-sided, and $p < 0.05$ was considered statistically significant. Expression data and network analyses were performed with the MeV and Cytoscape software suites[74,75]. Class comparison analyses were performed using BRB-ArrayTools v4.2.1.

References

1. Malvezzi, M. *et al*. European cancer mortality predictions for the year 2011. *Ann Oncol* **22**, 947–956, doi:10.1093/annonc/mdq774 (2011).
2. Espinosa, E. *et al*. The present and future of gene profiling in breast cancer. *Cancer Metastasis Rev* **31**, 41–46, doi:10.1007/s10555-011-9327-7 (2012).
3. Perou, C. M. *et al*. Molecular portraits of human breast tumours. *Nature* **406**, 747–752, doi:10.1038/35021093 (2000).
4. Parker, J. S. *et al*. Supervised risk predictor of breast cancer based on intrinsic subtypes. *J Clin Oncol* **27**, 1160–1167, doi:10.1200/JCO.2008.18.1370 (2009).
5. Prat, A., Ellis, M. J. & Perou, C. M. Practical implications of gene-expression-based assays for breast oncologists. *Nat Rev Clin Oncol* **9**, 48–57, doi:10.1038/nrclinonc.2011.178 (2012).
6. Hanash, S. Disease proteomics. *Nature* **422**, 226–232, doi:10.1038/nature01514 (2003).
7. Marko-Varga, G. *et al*. Personalized medicine and proteomics: lessons from non-small cell lung cancer. *J Proteome Res* **6**, 2925–2935, doi:10.1021/pr070046s (2007).
8. Pastwa, E., Somiari, S. B., Czyz, M. & Somiari, R. I. Proteomics in human cancer research. *Proteomics Clin Appl* **1**, 4–17 (2007).
9. Rifai, N., Gillette, M. A. & Carr, S. A. Protein biomarker discovery and validation: the long and uncertain path to clinical utility. *Nat Biotechnol* **24**, 971–983, doi:10.1038/nbt1235 (2006).
10. Gamez-Pozo, A. *et al*. Combined label-free quantitative proteomics and microRNA expression analysis of breast cancer unravel molecular differences with clinical implications. *Cancer Research*, doi:10.1158/0008-5472.CAN-14-1937 (2015).
11. Colijn, C. *et al*. Interpreting expression data with metabolic flux models: predicting Mycobacterium tuberculosis mycolic acid production. *PLoS Comput Biol* **5**, e1000489, doi:10.1371/journal.pcbi.1000489 (2009).

12. Picotti, P., Bodenmiller, B., Mueller, L. N., Domon, B. & Aebersold, R. Full Dynamic Range Proteome Analysis of S. cerevisiae by Targeted Proteomics. *Cell* **138**, 795–806, doi:10.1016/j.cell.2009.05.051 (2009).
13. Guedj, M. *et al.* A refined molecular taxonomy of breast cancer. *Oncogene* **31**, 1196–1206, doi:10.1038/onc.2011.301 (2012).
14. Miller, L. D. *et al.* An iron regulatory gene signature predicts outcome in breast cancer. *Cancer Res* **71**, 6728–6737, doi:10.1158/0008-5472.CAN-11-1870 (2011).
15. van de Vijver, M. J. *et al.* A gene-expression signature as a predictor of survival in breast cancer. *N Engl J Med* **347**, 1999–2009, doi:10.1056/NEJMoa021967 (2002).
16. Paik, S. *et al.* A multigene assay to predict recurrence of tamoxifen-treated, node-negative breast cancer. *N Engl J Med* **351**, 2817–2826, doi:10.1056/NEJMoa041588 (2004).
17. Sanchez-Navarro, I. *et al.* An 8-gene qRT-PCR-based gene expression score that has prognostic value in early breast cancer. *BMC Cancer* **10**, 336, doi:10.1186/1471-2407-10-336 (2010).
18. Troup, S. *et al.* Reduced expression of the small leucine-rich proteoglycans, lumican, and decorin is associated with poor outcome in node-negative invasive breast cancer. *Clin Cancer Res* **9**, 207–214 (2003).
19. Cawthorn, T. R. *et al.* Proteomic analyses reveal high expression of decorin and endoplasmin (HSP90B1) are associated with breast cancer metastasis and decreased survival. *PloS one* **7**, e30992, doi:10.1371/journal.pone.0030992 (2012).
20. Mefford, D. & Mefford, J. Stromal genes add prognostic information to proliferation and histoclinical markers: a basis for the next generation of breast cancer gene signatures. *PloS one* **7**, e37646, doi:10.1371/journal.pone.0037646 (2012).
21. Muraoka, S. *et al.* Strategy for SRM-based verification of biomarker candidates discovered by iTRAQ method in limited breast cancer tissue samples. *Journal of proteome research* **11**, 4201–4210, doi:10.1021/pr300322q (2012).
22. Cittelly, D. M. *et al.* Downregulation of miR-342 is associated with tamoxifen resistant breast tumors. *Mol Cancer* **9**, 317, doi:10.1186/1476-4598-9-317 (2010).
23. Miller, T. E. *et al.* MicroRNA-221/222 confers tamoxifen resistance in breast cancer by targeting p27Kip1. *J Biol Chem* **283**, 29897–29903, doi:10.1074/jbc.M804612200 (2008).
24. He, Y. J. *et al.* miR-342 is associated with estrogen receptor-alpha expression and response to tamoxifen in breast cancer. *Exp Ther Med* **5**, 813–818, doi:10.3892/etm.2013.915 (2013).
25. Perez-Rivas, L. G. *et al.* A microRNA signature associated with early recurrence in breast cancer. *PLoS One* **9**, e91884, doi:10.1371/journal.pone.0091884 (2014).
26. Dabiri, S. *et al.* The presence of stromal mast cells identifies a subset of invasive breast cancers with a favorable prognosis. *Mod Pathol* **17**, 690–695, doi:10.1038/modpathol.3800094 (2004).
27. Rajput, A. B. *et al.* Stromal mast cells in invasive breast cancer are a marker of favourable prognosis: a study of 4,444 cases. *Breast Cancer Res Treat* **107**, 249–257, doi:10.1007/s10549-007-9546-3 (2008).
28. Amini, R. M. *et al.* Mast cells and eosinophils in invasive breast carcinoma. *BMC cancer* **7**, 165, doi:10.1186/1471-2407-7-165 (2007).
29. della Rovere, F. *et al.* Mast cells in invasive ductal breast cancer: different behavior in high and minimum hormone-receptive cancers. *Anticancer Res* **27**, 2465–2471 (2007).
30. Baniwal, S. K., Chimge, N. O., Jordan, V. C., Tripathy, D. & Frenkel, B. Prolactin-induced protein (PIP) regulates proliferation of luminal A type breast cancer cells in an estrogen-independent manner. *PloS one* **8**, e62361, doi:10.1371/journal.pone.0062361 (2014).
31. Darb-Esfahani, S. *et al.* Gross cystic disease fluid protein 15 (GCDFP-15) expression in breast cancer subtypes. *BMC cancer* **14**, 546, doi:10.1186/1471-2407-14-546 (2014).
32. Luo, M. H. *et al.* Expression of mammaglobin and gross cystic disease fluid protein-15 in breast carcinomas. *Hum Pathol* **44**, 1241–1250, doi:10.1016/j.humpath.2012.10.009 (2013).
33. Parris, T. Z. *et al.* Clinical implications of gene dosage and gene expression patterns in diploid breast carcinoma. *Clin Cancer Res* **16**, 3860–3874, doi:10.1158/1078-0432.CCR-10-0889 (2010).
34. Parris, T. Z. *et al.* Additive effect of the AZGP1, PIP, S100A8 and UBE2C molecular biomarkers improves outcome prediction in breast carcinoma. *Int J Cancer* **134**, 1617–1629, doi:10.1002/ijc.28497 (2014).
35. Jablonska, K. *et al.* Prolactin-induced protein as a potential therapy response marker of adjuvant chemotherapy in breast cancer patients. *American journal of cancer research* **6**, 878–893 (2016).
36. Naderi, A. & Meyer, M. Prolactin-induced protein mediates cell invasion and regulates integrin signaling in estrogen receptor-negative breast cancer. *Breast cancer research: BCR* **14**, R111, doi:10.1186/bcr3232 (2012).
37. Naderi, A. & Vanneste, M. Prolactin-induced protein is required for cell cycle progression in breast cancer. *Neoplasia* **16**(329–342), e321–314, doi:10.1016/j.neo.2014.04.001 (2014).
38. Lehmann-Che, J. *et al.* Molecular apocrine breast cancers are aggressive estrogen receptor negative tumors overexpressing either HER2 or GCDFP15. *Breast cancer research: BCR* **15**, R37, doi:10.1186/bcr3421 (2013).
39. Johansson, H. J. *et al.* Proteomics profiling identify CAPS as a potential predictive marker of tamoxifen resistance in estrogen receptor positive breast cancer. *Clin Proteomics* **12**, 8, doi:10.1186/s12014-015-9080-y (2015).
40. Hernandez, V. J. *et al.* Cavin-3 dictates the balance between ERK and Akt signaling. *Elife* **2**, e00905, doi:10.7554/eLife.00905 (2013).
41. Wikman, H. *et al.* Clinical relevance of loss of 11p15 in primary and metastatic breast cancer: association with loss of PRKCDBP expression in brain metastases. *PloS one* **7**, e47537, doi:10.1371/journal.pone.0047537 (2012).
42. Cheng, Q. *et al.* Amplification and high-level expression of heat shock protein 90 marks aggressive phenotypes of human epidermal growth factor receptor 2 negative breast cancer. *Breast cancer research: BCR* **14**, R62, doi:10.1186/bcr3168 (2012).
43. Pick, E. *et al.* High HSP90 expression is associated with decreased survival in breast cancer. *Cancer research* **67**, 2932–2937, doi:10.1158/0008-5472.CAN-06-4511 (2007).
44. Echeverria, P. C., Bernthaler, A., Dupuis, P., Mayer, B. & Picard, D. An interaction network predicted from public data as a discovery tool: application to the Hsp90 molecular chaperone machine. *PloS one* **6**, e26044, doi:10.1371/journal.pone.0026044 (2011).
45. Scheufler, C. *et al.* Structure of TPR domain-peptide complexes: critical elements in the assembly of the Hsp70-Hsp90 multichaperone machine. *Cell* **101**, 199–210, doi:10.1016/S0092-8674(00)80830-2 (2000).
46. Meric-Bernstam, F. *et al.* Aberrations in translational regulation are associated with poor prognosis in hormone receptor-positive breast cancer. *Breast cancer research: BCR* **14**, R138, doi:10.1186/bcr3343 (2012).
47. DeBerardinis, R. J. *et al.* Beyond aerobic glycolysis: transformed cells can engage in glutamine metabolism that exceeds the requirement for protein and nucleotide synthesis. *Proc Natl Acad Sci USA* **104**, 19345–19350, doi:10.1073/pnas.0709747104 (2007).
48. Krishnan, K. *et al.* miR-139-5p is a regulator of metastatic pathways in breast cancer. *Rna* **19**, 1767–1780, doi:10.1261/rna.042143.113 (2013).
49. Lowery, A. J. *et al.* MicroRNA signatures predict oestrogen receptor, progesterone receptor and HER2/neu receptor status in breast cancer. *Breast Cancer Res* **11**, R27, doi:10.1186/bcr2257 (2009).
50. Gyorffy, B. *et al.* Multigene prognostic tests in breast cancer: past, present, future. *Breast Cancer Res* **17**, 11, doi:10.1186/s13058-015-0514-2 (2015).
51. Gamez-Pozo, A. *et al.* Shotgun proteomics of archival triple-negative breast cancer samples. *Proteomics Clin Appl* **7**, 283–291, doi:10.1002/prca.201200048 (2013).
52. Gamez-Pozo, A. *et al.* Protein phosphorylation analysis in archival clinical cancer samples by shotgun and targeted proteomics approaches. *Mol Biosyst* **7**, 2368–2374, doi:10.1039/c1mb05113j (2011).

53. Cox, J. & Mann, M. MaxQuant enables high peptide identification rates, individualized p.p.b.-range mass accuracies and proteome-wide protein quantification. *Nat Biotechnol* **26**, 1367–1372, doi:10.1038/nbt.1511 (2008).
54. Cox, J. *et al*. Andromeda: a peptide search engine integrated into the MaxQuant environment. *J Proteome Res* **10**, 1794–1805, doi:10.1021/pr101065j (2011).
55. Deeb, S. J., D'Souza, R. C. J., Cox, J., Schmidt-Supprian, M. & Mann, M. Super-SILAC Allows Classification of Diffuse Large B-cell Lymphoma Subtypes by Their Protein Expression Profiles. *Molecular & Cellular Proteomics* **11**, 77–89, doi:10.1074/mcp.M111.015362 (2012).
56. Johnson, W. E., Li, C. & Rabinovic, A. Adjusting batch effects in microarray expression data using empirical Bayes methods. *Biostatistics* **8**, 118–127, doi:10.1093/biostatistics/kxj037 (2007).
57. Sanchez-Navarro, I. *et al*. Comparison of gene expression profiling by reverse transcription quantitative PCR between fresh frozen and formalin-fixed, paraffin-embedded breast cancer tissues. *Biotechniques* **48**, 389–397, doi:10.2144/000113388 (2010).
58. Tusher, V. G., Tibshirani, R. & Chu, G. Significance analysis of microarrays applied to the ionizing radiation response. *Proc Natl Acad Sci USA* **98**, 5116–5121, doi:10.1073/pnas.091062498 (2001).
59. Schwarz, G. Estimating the dimension of a model. *Annals of Statistics* **6**, 461–464 (1978).
60. R Core Team. (R Foundation for Statistical Computing, Vienna, Austria., 2013).
61. Abreu, G. C. G., Edwards, D. & Labouriau, R. High-Dimensional Graphical Model Search with the gRapHD R Package. *Journal of Statistical Software* **37**, 1–18 (2010).
62. Huang da, W., Sherman, B. T. & Lempicki, R. A. Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nat Protoc* **4**, 44–57, doi:10.1038/nprot.2008.211 (2009).
63. Huang da, W., Sherman, B. T. & Lempicki, R. A. Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists. *Nucleic Acids Res* **37**, 1–13, doi:10.1093/nar/gkn923 (2009).
64. Thiele, I. *et al*. A community-driven global reconstruction of human metabolism. *Nature biotechnology* **31**, 419–425, doi:10.1038/nbt.2488 (2013).
65. Orth, J. D., Thiele, I. & Palsson, B. O. What is flux balance analysis? *Nature biotechnology* **28**, 245–248, doi:10.1038/nbt.1614 (2010).
66. Barker, B. E. *et al*. A robust and efficient method for estimating enzyme complex abundance and metabolic flux from expression data. *Computational Biology and Chemistry* **59**(Part B), 98–112, doi:10.1016/j.compbiolchem.2015.08.002 (2015).
67. Schellenberger, J. *et al*. Quantitative prediction of cellular metabolism with constraint-based models: the COBRA Toolbox v2.0. *Nat Protoc* **6**, 1290–1307, doi:10.1038/nprot.2011.308 (2011).
68. Deutsch, E. W., Lam, H. & Aebersold, R. PeptideAtlas: a resource for target selection for emerging targeted proteomics workflows. *EMBO Rep* **9**, 429–434, doi:10.1038/embor.2008.56 (2008).
69. MacLean, B. *et al*. Skyline: an open source document editor for creating and analyzing targeted proteomics experiments. *Bioinformatics* **26**, 966–968, doi:10.1093/bioinformatics/btq054 (2010).
70. Wright, G. W. & Simon, R. M. A random variance model for detection of differential gene expression in small microarray experiments. *Bioinformatics* **19**, 2448–2455, doi:10.1093/bioinformatics/btg345 (2003).
71. Simon, R., Radmacher, M. D., Dobbin, K. & McShane, L. M. Pitfalls in the Use of DNA Microarray Data for Diagnostic and Prognostic Classification. *Journal of the National Cancer Institute* **95**, 14–18, doi:10.1093/jnci/95.1.14 (2003).
72. Fan, C. *et al*. Concordance among gene-expression-based predictors for breast cancer. *N Engl J Med* **355**, 560–569, doi:10.1056/NEJMoa052933 (2006).
73. Hu, Z. *et al*. The molecular portraits of breast tumors are conserved across microarray platforms. *BMC Genomics* **7**, 96, doi:10.1186/1471-2164-7-96 (2006).
74. Saeed, A. I. *et al*. TM4: a free, open-source system for microarray data management and analysis. *Biotechniques* **34**, 374–378 (2003).
75. Shannon, P. *et al*. Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome research* **13**, 2498–2504, doi:10.1101/gr.1239303 (2003).

Acknowledgements

The authors would like to acknowledge funding from grants PI12/00444, PI12/01016 and PI15/01310 from the Instituto de Salud Carlos III, Spanish Economy and Competitiveness Ministry, Spain, and co-funded by the FEDER program, "Una forma de hacer Europa". This study has also been supported by the PRIME-XS project, grant agreement number 262067, funded by the EU's Seventh Framework Program for Research. AG-P and RL-V are supported by Instituto de Salud Carlos III, and the Spanish Economy and Competitiveness Ministry grants, CA12/00258 and CA12/00264, respectively. We want to particularly acknowledge the patients in this study for their participation and to the IdiPAZ, I+12 and O+EHUN Biobanks for the generous gifts of clinical samples used in this study. LT-F is supported by the Spanish Economy and Competitiveness Ministry (DI-15-07614). IdiPAZ, I+12 and O+EHUN Biobanks are supported by Instituto de Salud Carlos III, Spanish Economy and Competitiveness Ministry (RD09/0076/00073, RD09/0076/00118 and RD09/0076/00140, respectively) and FarmaIndustria, through the Cooperation Program in Clinical and Translational Research of the Community of Madrid and Basque Autonomous Community.

Author Contributions

All the authors have directly participated in the preparation of this manuscript and have approved the final version submitted. They declare no ethical conflicts of interest. J.B.-S., R.L.-V. and A.G.-P. contributed the RNA and protein extraction. P.N., N.S. and J.G. contributed the mass spectrometry data. J.M.A., H.N. and P.M. contributed the probabilistic graphical models. M.D.-A. and F.G.M. contributed the GPR rule method. P.M.d.P., P.Z., J.F., E.C. and E.E. contributed the clinical data and the analyses related. A.G.-P., G.P.-V., A.Z.-M., and J.B.-S. contributed to the design of the study and the statistical and gene ontology analyses. A.G.-P. drafted the manuscript. L.T.-F. and R.G.-R. contributed the FBA analyses. J.A.F.V., P.M.d.P., P.Z., J.F., E.C. and E.E. conceived of the study and participated in its design and interpretation. J.A.F.V. coordinated the study. All the authors have read and approved the final manuscript.

Additional Information

Supplementary information accompanies this paper at doi:10.1038/s41598-017-10493-w

Competing Interests: J.A.F.V., E.E. and A.G.-P. are shareholders in Biomedica Molecular Medicine S.L. L.T.-F. is an employee of Biomedica Molecular Medicine S.L. The other authors declare that they have no competing interests.

www.impactjournals.com/oncotarget/ Oncotarget, 2018, Vol. 9, (No. 11), pp: 9645-9660

Research Paper

Molecular characterization of breast cancer cell response to metabolic drugs

Lucía Trilla-Fuertes[1,2], Angelo Gámez-Pozo[1,2], Jorge M. Arevalillo[3], Mariana Díaz-Almirón[4], Guillermo Prado-Vázquez[1], Andrea Zapater-Moros[1], Hilario Navarro[3], Rosa Aras-López[5], Irene Dapía[6,7], Rocío López-Vacas[1], Paolo Nanni[8], Sara Llorente-Armijo[1], Pedro Arias[6,7], Alberto M. Borobia[9], Paloma Maín[10], Jaime Feliú[11,12,13], Enrique Espinosa[11,12] and Juan Ángel Fresno Vara[1,2,12]

[1]Molecular Oncology and Pathology Lab, Institute of Medical and Molecular Genetics-INGEMM, La Paz University Hospital-IdiPAZ, Madrid, Spain

[2]Biomedica Molecular Medicine SL, Madrid, Spain

[3]Operational Research and Numerical Analysis, National Distance Education University (UNED), Madrid, Spain

[4]Biostatistics Unit, La Paz University Hospital-IdiPAZ, Madrid, Spain

[5]Congenital Malformations Lab, Institute of Medical and Molecular Genetics-INGEMM, La Paz University Hospital, IdiPAZ, Madrid, Spain

[6]Pharmacogenetics Lab, Institute of Medical and Molecular Genetics-INGEMM, La Paz University Hospital-IdiPAZ, Autonomous University of Madrid, Madrid, Spain

[7]Biomedical Research Networking Center on Rare Diseases-CIBERER, ISCIII, Madrid, Spain

[8]Functional Genomics Center Zurich, University of Zurich/ETH Zurich, Zurich, Switzerland

[9]Clinical Pharmacology Department, La Paz University Hospital School of Medicine, IdiPAZ, Autonomous University of Madrid, Madrid, Spain

[10]Department of Statistics and Operations Research, Faculty of Mathematics, Complutense University of Madrid, Madrid, Spain

[11]Medical Oncology Service, La Paz University Hospital-IdiPAZ, Madrid, Spain

[12]Biomedical Research Networking Center on Oncology-CIBERONC, ISCIII, Madrid, Spain

[13]Cátedra UAM-AMGEN, Universidad Autónoma de Madrid, Madrid, Spain

Correspondence to: *Juan Ángel Fresno Vara,* ***email:*** *juanangel.fresno@salud.madrid.org*

Keywords: *breast cancer; flux balance analysis; metabolism; perturbation experiments; proteomics*

Received: *October 29, 2017* ***Accepted:*** *January 03, 2018* ***Published:*** *January 08, 2018*

ABSTRACT

Metabolic reprogramming is a hallmark of cancer. It has been described that breast cancer subtypes present metabolism differences and this fact enables the possibility of using metabolic inhibitors as targeted drugs in specific scenarios. In this study, breast cancer cell lines were treated with metformin and rapamycin, showing a heterogeneous response to treatment and leading to cell cycle disruption. The genetic causes and molecular effects of this differential response were characterized by means of SNP genotyping and mass spectrometry-based proteomics. Protein expression was analyzed using probabilistic graphical models, showing that treatments elicit various responses in some biological processes such as transcription. Moreover, flux balance analysis using protein expression values showed that predicted growth rates were comparable with cell viability measurements and suggesting an increase in reactive oxygen species response enzymes due to metformin treatment. In addition, a method to assess flux differences in whole pathways was proposed. Our results show that these diverse

approaches provide complementary information and allow us to suggest hypotheses about the response to drugs that target metabolism and their mechanisms of action.

INTRODUCTION

Breast cancer is one of the most prevalent cancers in the world [1]. In clinical practice, breast cancer is divided according to three biomarkers, estrogen receptor (ER), progesterone receptor (PR) and Her2; into positive hormonal receptors (ER+), HER2+ and triple negative (TNBC), characterized by a lack of expression of these receptors. These biomarkers are associated with specific treatments. ER+ tumors are treated with selective ER modulator or aromatase inhibitors [2, 3] and Her2 tumors are treated with antibodies against this receptor [4]. However, TNBC tumors don´t have a specific treatment. In addition to the clinical classification, molecular profiles based on mRNA expression are also established [5].

Reprogramming of cellular metabolism is a hallmark of cancer [6]. Normal cells obtain energy mainly from mitochondrial metabolism, but cancer cells show increased glucose uptake and fermentation into lactate, which is known as the Warburg effect or aerobic glycolysis [7]. Cancer cells also exhibit increased glutamine uptake to maintain the pool of nonessential amino acids and to further increase lactate production [8]. In addition, we previously observed significant differences in glucose metabolism between two of the main breast cancer subtypes: ER+ and TNBC [9, 10].

Metabolic alterations enable the possibility of using metabolic inhibitors as targeted drugs. Metformin (MTF), a drug for diabetes, has begun clinical trials in cancer patients [11]. It activates AMP-activated protein kinase and subsequently inhibits mammalian target of rapamycin (mTOR) [12]. On the other hand, everolimus, a rapamycin analog, has clinical activity and has been approved for use in patients with breast cancer and other tumors [13]. Rapamycin (RP) or sirolimus was the first available mammalian target of rapamycin (mTOR) inhibitor.

High-throughput mass spectrometry-based proteomics allow the quantification of thousands of proteins and the acquisition of direct information about biological process effectors. Combined with probabilistic graphical models (PGM), proteomics enables the characterization of various biological processes between different conditions using expression data without other *a priori* information [9, 10].

Flux Balance Analysis (FBA) is a widely used approach for modeling biochemical and metabolic networks in a genome scale [14–16]. FBA calculates the flow of metabolites through metabolic networks, allowing the prediction of growth rates or the rate of production of a metabolite. It has traditionally been used to estimate microorganism growth rates [17]. However, with the appearance of complete reconstructions of human metabolism, FBA has been applied to other areas such as the modelling of red blood cells metabolism [18] or the study of the Warburg effect in cancer cell lines [19].

In the present study, we used proteomics and computational methods, such as PGM and a genome-scale model of metabolism analyzed using FBA, to explore the molecular consequences of metformin and rapamycin treatment in breast cancer cell lines.

RESULTS

Design of the study

We studied response against MTF and RP in six breast cancer cell lines, establishing sub-lethal doses to perform subsequent perturbation experiments. On the other hand, we studied single nucleotide polymorphisms (SNP) to check if the heterogeneity to treatment response observed among breast cancer cell lines can be associated to genetic causes. Then, perturbation experiments followed by mass spectrometry-based proteomics were done to characterize these differences at the molecular level. Differential protein expression patterns were analyzed and probabilistic graphical models (PGM) and flux balance analysis (FBA) were performed in order to characterize the molecular consequences of response against MTF and RP (Figure 1). SNP genotyping was used to study genetic variants associated with response and proteomics data were used to complement this information, study functional differences by probabilistic graphical models and improve prediction accuracy of FBA. PGM allowed characterizing differences due to the treatments at functional level and FBA was useful to study effects in the metabolic pathways. These approaches provide complementary information about genetic causes and molecular effects respectively.

Breast cancer cell lines showed heterogeneous response when treated with drugs against metabolic targets

First, we evaluated the response of ER+ and TNBC breast cancer cell lines treated with two drugs targeting metabolism, metformin (MTF) and rapamycin (RP). Cell viability was assessed for six breast cancer cell lines, three ER+ (T47D, MCF7 and CAMA1) and three TNBC (MDAMB231, MDAMB468 and HCC1143). Dose-response curves for each drug treatment in each cell were calculated (Tables 1 and 2). A heterogeneous response was observed among breast cancer cell lines treated with a range of MTF and RP concentrations (Figure 2). Regarding RP, this heterogeneous response is related to breast cancer subtypes, showing an increased effect over ER+ cell line viability compared with those of TNBC.

SNP genotyping of breast cancer cell lines

SNP genotyping was performed to evaluate the association of polymorphisms to MTF and RP treatment response. Polymorphisms previously related to these drugs sensitivity were studied using a custom expression array. Regarding the response to MTF, polymorphism rs2282143 in *SLC22A1* was detected in homozygosis in MDAMB468 cells. This SNP appears with a frequency of 8% in the black population, which is the population origin of this cell line, and it is associated with decreased clearance of MTF. On the other hand, the rs628031 polymorphism, also in *SLC22A1*, was found in homozygosis in MCF7 and HCC1143 cells and in heterozygosis with a possible duplication in MDAMB468 cells. The presence of this polymorphism has been associated with a decreased response to MTF (PharmGKB; www.pharmgkb.org) (Supplementary Table 1).

Regarding the response to RP, MDAMB468 cells present a polymorphism in heterozygosis in *CYP3A4* (rs2740574), which has been previously related to a requirement for an increased dose of RP as compared with a wild-type homozygote (PharmGKB; www.pharmgkb.org). Additionally, rs2868177 SNP in *POR* gene was detected in heterozygosis in hormone receptor-positive cell lines. The relationship of rs2868177 with RP or another rapalog has not been previously described, although it is demonstrated that POR regulates *CYP3A* family [20]. On the other hand, rs1045642 SNP in *ABCB1* gene appears in heterozygosis in all ER+ cell lines, but its effect regarding RP concentration is controversial (PharmGKB; www.pharmgkb.org) (Supplementary Table 1).

Molecular characterization of breast cancer cell lines response to treatment with drugs against metabolic targets using perturbation experiments and proteomics

SNP genotyping did not fully explain the heterogeneous response between cell lines to MTF and RP treatment, thus we characterized the molecular basis of this heterogeneous response using proteomics in a perturbation experimental setting. Six breast cancer cell lines, treated or not with suboptimal concentrations of MTF and RP (40 mM of MTF [except for MDAMB468, in which a 20 mM concentration was used] and 625 nM of RP) were analyzed in duplicate using shotgun proteomics. Raw data normalization was performed adjusting by duplicate values as previously described [9]. Mass spectrometry-based proteomics allowed the detection of 4052 proteins presenting at least two unique peptides and detectable expression in at least 75% of the samples (Supplementary Table 2). No decoy protein passed through these additional filters. Label-free quantification values from these 4052 proteins were used in subsequent analyses.

We first identified proteins with differential expression between the treated and the control cells. Proteins with delta expression values between the control and treated cells higher than 1.5 or lower than −1.5 were identified for each cell line/treatment combination (Supplementary Tables 3 and 4). Then, gene ontology analyses of either increased or decreased proteins was performed. Regarding MTF treatment, MCF7 cells showed decreased expression of proteins related to mitochondria and cell cycle and increased expression of proteins involved in mitochondria and cytoskeleton as majority ontologies. T47D cells presented increased expression of proteins mostly related to mitochondria and the Golgi apparatus. CAMA1 proteins showing differential expression did not shown overrepresented functions. MDAMB231 cells showed decreased expression of proteins mostly related to mitochondria. MDAMB468 cells presented decreased expression of proteins also related to mitochondria, and increased expression in proteins mainly related to the extracellular matrix. Finally, HCC1143 showed decreased expression in proteins, mostly related to mitochondria and mRNA processing, and increased expression in proteins related to cytosol and protein binding.

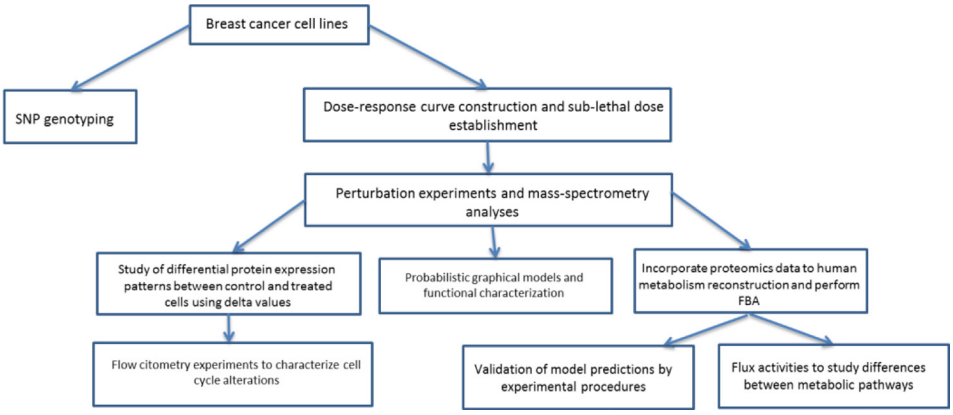


Figure 1: Workflow followed in this study.

Table 1: Cell viability measurements in MTF treated cells

MTF mM	0	5	10	20	40	80	160
MCF7	100.00	135.07	95.00	61.49	30.61	28.36	2.47
T47D	100.00	85.74	70.15	59.87	42.11	7.10	0.00
CAMA1	100.00	88.08	112.76	93.70	108.67	63.25	3.49
MDAMB231	100.00	65.08	58.36	57.78	37.82	11.45	1.77
MDAMB468	100.00	40.05	55.39	21.82	1.31	1.71	0.00
HCC1143	100.00	105.48	85.25	73.19	52.89	20.49	0.00

Cell viability measurements in six breast cancer cell lines treated with MTF (0–160 mM). Red-white-blue color scale.

Table 2: Cell viability measurements in RP treated cells

RP nM	0	156.25	312.5	625	1250	2500	5000	10000
MCF7	100.00	29.36	22.34	31.62	19.88	16.29	7.53	3.32
T47D	100.00	33.02	33.76	43.74	24.39	17.73	8.69	11.15
CAMA1	100.00	70.22	46.25	45.99	26.28	22.46	13.45	7.71
MDAMB231	100.00	79.92	82.09	67.84	62.16	62.43	31.95	24.50
MDAMB468	100.00	48.25	48.51	71.92	75.75	52.74	55.31	4.49
HCC1143	100.00	125.74	136.39	137.53	144.66	130.58	85.55	24.85

Cell viability measurements in six breast cancer cell lines treated with RP (0–10,000 nM). Red-white-blue color scale.

Differentially expressed proteins were compared with gene interaction information contained in the Comparative Toxicogenomics Database. PIR, RELA, SIRT5, CMBL, PPP4R2 and MYD88 showed decreased expression, whereas SIRT2, SERPINE1 and HTATIP2 proteins showed increased expression in cells treated with MTF in both the database and in our experiments in at least one cell line.

Concerning RP treatment, MCF7 showed decreased expression in proteins mainly related to cellular transport and an increased expression in proteins related to the mitochondrial matrix. T47D presented decreased expression in proteins involved in cell division and an increase in proteins related to lysosomes. CAMA1 had a decrease in expression of proteins associated with mRNA processing, splicing and mitochondria and an increase in the expression of proteins related to mitochondria, apoptosis processes and especially with the role of mitochondria in the apoptotic pathway. MDAMB231 had a decrease in proteins related to mRNA processing and cytoskeleton and an increase in proteins related to exosomes. MDAMB468 proteins showing differential expression did not shown overrepresented functions. Lastly, HCC1143 showed a decreased expression in proteins related to lysosomes and an increased expression in proteins related to mitochondria.

Gene interaction information contained in the Comparative Toxicogenomics Database showed a decrease in CDK4, CKS1B, COL1A1, IGFBP5, KIFC1, mTOR and SCD expression and an increase in CASP8, NR3C1, PKP4, RPS27L, TEAD1 and XIAP due to RP treatment in both the database and in our experiments in at least one cell line.

Then, we applied linear regression models using protein expression data to discover molecular markers predicting the response to MTF and RP treatment. MMGT1, IDH1, PSPC1 and TACO1 showed the strongest correlation with the response to MTF (Supplementary Table 5), whereas ACADSB, CCD58, MPZL1 and SBSN correlated with the response to RP (Supplementary Table 6).

The next step was to explore molecular functions and biological pathways deregulated by MTF and RP treatment. Protein expression data from treated and untreated cells were used to build a probabilistic graphical model without other *a priori* information. The resulting graph was processed to seek a functional structure (Figure 3), i.e., whether the proteins included in each branch of the tree had some relationship regarding their function, as previously described [9]. Thus, we divided our graph into 36 branches and performed gene ontology analyses. Twenty-nine of them had a significant enrichment in proteins related to a specific biological function.

Functional node activity was calculated for each branch with a defined biological function using protein delta values between control and treated cells. MTF treatment caused decreased activity in mitochondria B, mRNA processing, DNA replication and ATP binding functional nodes in all cell lines (Supplementary Figure 1). In the case of RP treatment, decreased activity was

observed in mRNA processing node activity in all cell lines (Supplementary Figure 2).

Functional node activities were then evaluated using multiple linear regression models to explore the relationship between functional deregulation and MTF/RP treatment. The response to RP treatment was explained using metabolism A and B node activities (adjusted $R^2 = 0.955$). Metabolism A node is primarily related to fatty acid biosynthesis and pyrimidine metabolism and Metabolism B node is related to glycolysis, oxidative phosphorylation and carbon metabolism (Supplementary Table 7). The response to MTF could not be predicted using this approach.

Cytometry experiments showed cytostatic effects of metformin and rapamycin treatment in breast cancer cells

The proteomics analysis workflow and gene ontology of delta values suggested that MTF and RP cause cell cycle alterations due to the recurrent replay of cell cycle category in ontology analyses. To confirm this hypothesis, flow cytometry assessment of the cell cycle was performed. MCF7 and MDAMB231 cells treated with MTF showed an increased proportion of G2/M cells when compared with the control, suggesting a cell cycle arrest in the G2 phase. However, CAMA1 cells show an increase in G1 phase percentage. Regarding RP, the ER+ cell lines MCF7 and T47D treated with RP presented an increased percentage of G0/G1 cells when compared with the control, suggesting a cell cycle arrest in G1. On the other hand, the HCC1143 cycle showed an increase in G2 percentages (Figure 4, Supplementary Table 8).

Flux balance analysis predicts alterations in growth rate in metformin-treated cells

To evaluate the impact of MTF and RP treatment on cellular metabolism, an FBA, including proteomics data from perturbation experiments, was applied to estimate cell growth rates for both control and treatment conditions. FBA can be used to evaluate a metabolic computational model to obtain a prediction of the tumor growth rate. This analysis can incorporate gene or protein expression data to improve prediction accuracy. Protein data allows constraining 2414 reactions of the 4253 reactions contained in Recon2, which have a defined gene-protein-reaction (GPR) rule, which include information of the genes/proteins involved in each enzymatic reaction. FBA predicts a lower growth rate in TNBC and MCF7 cell lines treated with MTF compared with control cells. However, it predicts a higher growth rate in the case of CAMA1 cells treated with MTF (Supplementary Table 9). FBA predicts no differences in growth rate between the control and the RP-treated cells.

FBA growth predictions match with experimental data from breast cancer cell cultures

Growth studies in ER+ (MCF7 and T47D) and TNBC (MDAMB231 and MDAMB468) cell lines were performed to validate FBA predictions using a dynamic FBA cell growth model. The starting concentration of glucose in medium (200 mg/dl) was incorporated into the dynamic FBA inputs. Initial experimental cell density was estimated by direct counting of seeded cells in the delimited area and used as a function input (MCF7= 37, T47D= 31, MDAMB231= 30 and MDAMB468 = 58 cells respectively). Growth rate predictions were comparable with experimental measurements in cell cultures over 72 hours (Figure 5). The highest deviation in absolute values is observed in MDAMB468 cells, whereas MCF7 predictions coincided with experimental observations.

Flux activity characterization

In order to compare fluxes from complete metabolic pathways between untreated and treated cell lines, a new

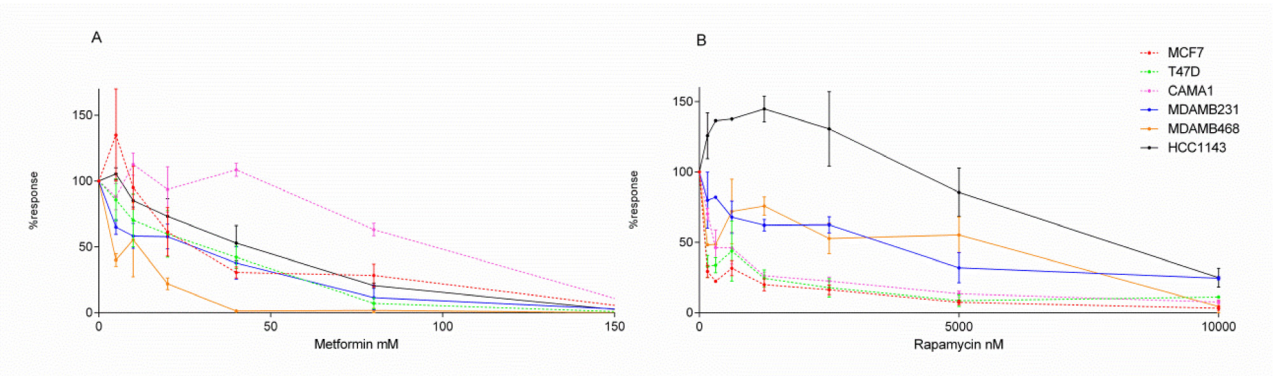


Figure 2: Dose-response curves. Dose-response curves of breast cancer cell lines treated with (**A**) MTF (0–160 mM) or (**B**) RP (0–10,000 nM). ER+ cell lines are represented as discontinuous lines and TNBC cells as continuous lines.

method named flux activities was proposed. Flux activities were calculated as the sum of the fluxes of each reaction in each pathway defined in the Recon2. Then, flux activities were used to build linear regression models to predict response. Pathways related to glutamate and pyruvate metabolism were related to response to MTF (adjusted R^2 =1) (Supplementary Table 10). In the case of RP, pathway fluxes that predict response against RP are cholesterol metabolism and valine, leucine and isoleucine metabolism (adjusted R^2 = 1) (Supplementary Table 11).

Flux analyses predict activation of ROS enzymes by metformin

With the aim of identifying reactions that changed as a consequence of treatment, we performed a Monte Carlo analysis and chose the solution with the maximum sum of fluxes because it was representative of protein data (i.e., if a protein was measured, it indicated that the protein must be used by the cell). After that, we applied flux variability analysis (FVA) to calculate the possible maximum and minimum fluxes for each reaction, and therefore, the range of fluxes for each reaction. Next, we selected reactions showing a flux change between the control and the treated cells over 95% of this range. As long as FBA provides a unique optimal tumor growth rate, multiple combinations of fluxes can lead to this optimal value. Therefore, we confirmed that the results from the maximum flux solution were consistent throughout the multiple-solution landscape using a Monte Carlo approach to study a range of representative flux solutions from all possible solutions that optimize the tumor growth rate. Of all the candidates evaluated, we would like to highlight that FBA predicts a null catalase flux in control cells with the exception of HCC1143 cells, showing constitutive catalase activation. In MDAMB231 and MCF7 cell lines treated with MTF, the model predicts an activation of this reaction, whereas CAMA1 cells showed no response to MTF treatment regarding catalase activation (Supplementary Figure 3, Supplementary Files 1–12).

Additionally, our model predicted that superoxide dismutase (SPODM) fluxes were increased in MCF7 and HCC1143 cell lines treated with MTF, but not in MDAMB231 cells. Predictions for CAMA1 cells showed high SPODM fluxes in both control and MTF treated cells (Supplementary Figure 4 and Supplementary Files 1–12).

Finally, the Monte Carlo approach predicted an increase in nitric oxide synthase flux and, as a consequence, an increase in nitric oxide (NO) production due to MTF treatment (Supplementary Figure 5).

On the other hand, proteomics data showed an increased expression of catalase in cells treated with MTF, with the exception of the CAMA1 cell line (Supplementary Table 8). It also showed an increased

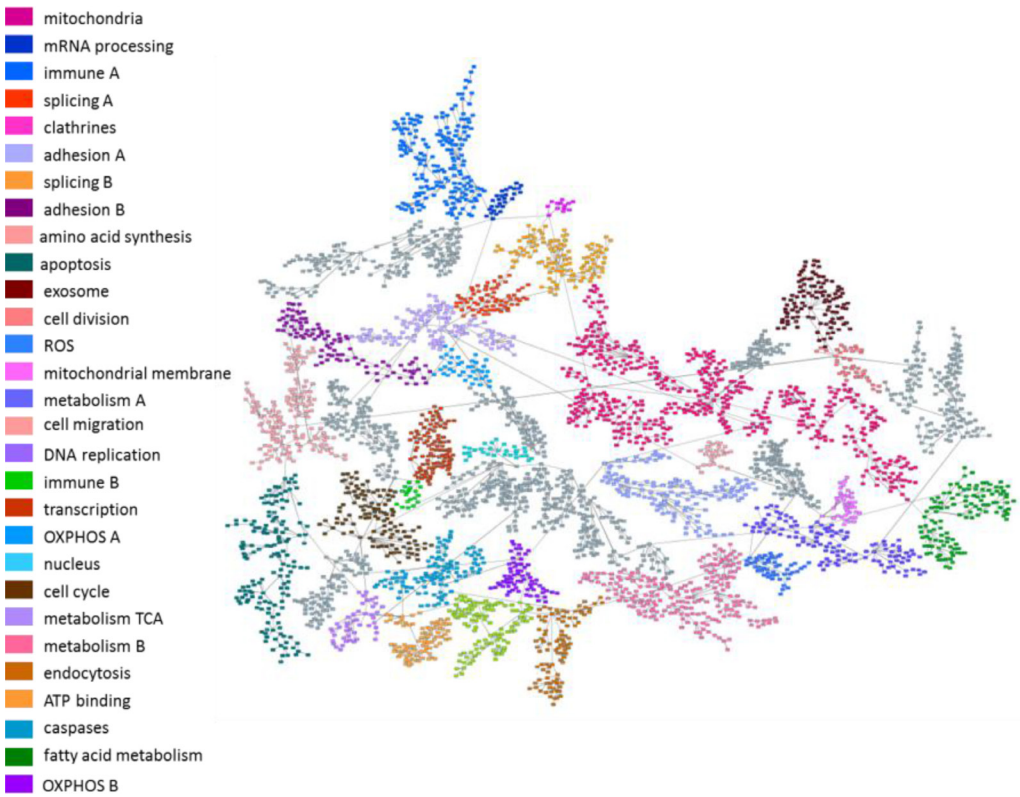


Figure 3: Probabilistic graphical model. Probabilistic graphical model using protein expression data of control and treated breast cancer cell lines. Gray nodes lack a specific function.

expression of SPODM in cells treated with MTF, although, SPODM expression was generally lower in MDAMB231 cells than in the rest of the cell lines (Supplementary Table 12). No protein expression data from NO were obtained.

Superoxide dismutase measurements confirm superoxide dismutase activation predictions

SPODM activities were measured in the control and in the MTF-treated cells using an enzyme activity

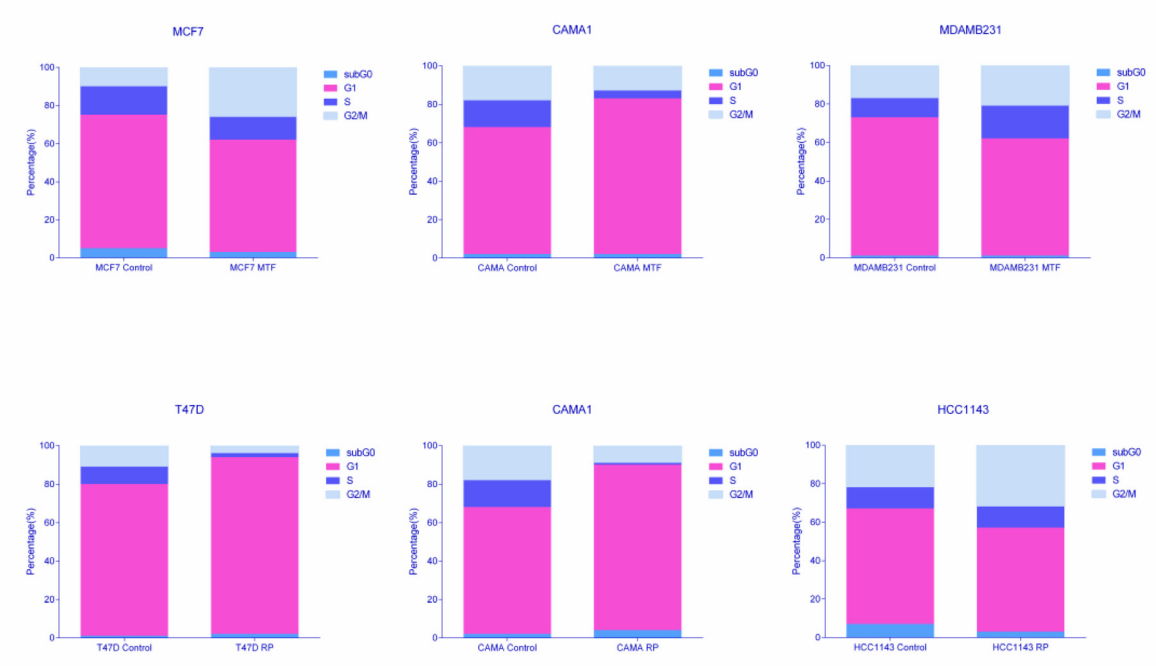


Figure 4: Percentages of cells in each cell cycle phase obtained by flow cytometry analyses.

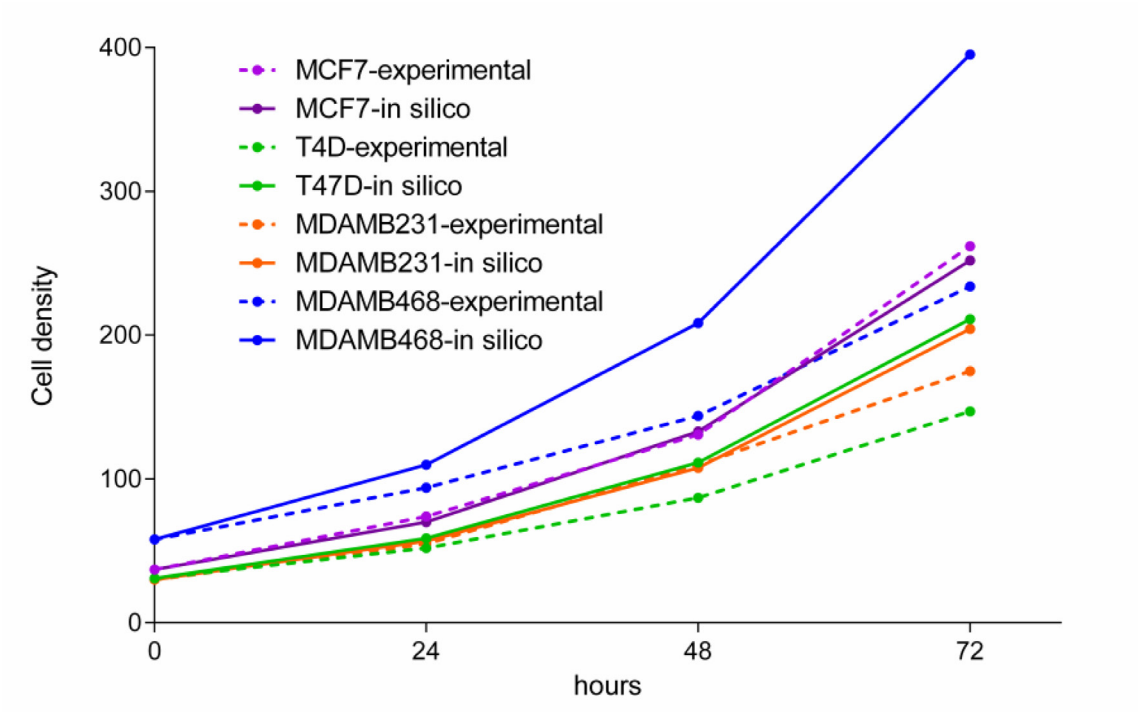


Figure 5: Experimental measurements of cell growth over 72 hours and a model simulation of growth during the same time period.

assay. With the exception of the MCF7 cell line, model predictions were confirmed. In HCC1143, SPODM activity is similar between the control and the treated cells. On the other hand, MDAMB231 had the lowest SPODM activity, as shown in model predictions, and CAMA1 cells had the highest SPODM activity in the control and in the MTF-treated cells, as predicted in the model (Table 3).

DISCUSSION

In this study, drugs targeting metabolism elicited changes related to cell cycle and oxidative stress in breast cancer cell lines. A high-throughput proteomics approach, coupled with a metabolism computational model, was useful to predict most of these changes and propose new mechanisms of action and effects of these drugs. To our knowledge, this is the first study that combines proteomics data with this type of computational analyses to study drug's mechanism of actions in breast cancer. However, FBA was successfully used in ovarian cancer cells to propose new therapeutic targets and to study the effect of drugs targeting metabolism and their synergies [21].

In previous studies, we observed significant differences between ER+ and TNBC glucose metabolism, which showed lactate production to be higher in TNBC cells than in ER+ cells [9]. These metabolic alterations suggest the possibility of using drugs against metabolic targets in patients with breast cancer.

Our results show that breast cancer cells' response to drugs targeting metabolism is heterogeneous. MTF treatment showed a broad effect on cell proliferation, with CAMA1 cells being the most resistant to this treatment. In the case of RP, the response depends on breast cancer subtype; it is effective in ER+ cell lines but not in those of TNBCs, resembling clinical results (a derivative of RP is used in women with hormone-receptor-positive breast cancer) [13].

With the aim of studying polymorphisms that could explain this heterogeneous cell response, an SNP array was used. Therefore, the high sensitivity to MTF showed by MDAMB468 cells could be partly due to rs2282143 SNP in the *SLC22A1* carrier, which is related to decreased clearance of MTF. In addition, *SLC22A1* rs628031, previously associated with a poorest response against MTF, was presented as homozygotic in the MCF7 and HCC1143 cell lines. ER+ cell lines presented heterozygosis in the *ABCB1* rs1045642 polymorphism, although the effects of this polymorphism in RP treatment response are not yet clear. In *CYP3A4*, rs2740574, which is related to higher requirement of sirolimus, is shown as heterozygotic in the MDAMB468 cell line.

We discovered several differences between the MTF-treated cells and the control cells. Some of these differential proteins identified matched with described interactions in the Comparative Toxicogenomics Database, such as increased expression of SIRT2 and HTATIP2 and decreased expression of SIRT5, PPP4R2 and MYD88 proteins due to MTF treatment. Increased SIRT2 protein expression induced by MTF treatment has been previously described [22]. SIRT2 also enhances gluconeogenesis, plays an important inhibitory role in inflammation and elevates ROS defense [23]. The effect of increased ROS stress response complies with our model predictions. Moreover, MTF treatment results in decreased SIRT5 expression [22]. This decrease is also related to differences observed in flux predictions between treated and control cells. It has been reported that SIRT5 is involved in the regulation of SPODM 1 activity [24], in accordance with our FBA prediction of SPODM activation in response to ROS stress in cells treated with MTF. On the other hand, TACO1, PSPC1, IDH1 and MMGT1 protein expression predict response to MTF treatment. *IDH1* mutations were previously related to hypersensitivity to biguanides [25]. PGM have shown that MTF treatment caused a decreased node activity in mRNA processing, DNA replication, mitochondria B and ATP binding nodes.

We also found several differences concerning RP treatment, such as an increased expression of NR3C1 and RPS27L proteins and a decreased expression of CKS1B, COL1A1, IGFBP5, SCD, mTOR and CDK4 proteins, as previously reported [26]. CDK4/6 inhibition robustly suppressed cell cycle progression of ER+/HER2- cellular models and complements the activity of limiting estrogen [27]. RP treatment also results in decreased expression of CKS1B mRNA [28]. Knockdown of CKS1 expression promotes apoptosis of breast cancer cells [29]. RP decreased expression of KIFC1 mRNA [30], whose overexpression is pro-proliferative [31]. RP treatment also results in increased activity of the NR3C1 protein [32]. NR3C1 encodes the glucocorticoid receptor, which is involved in the inflammation response and which has an anti-proliferative effect [33]. RP enhances TP73 binding to the RPS27L promoter, a direct p53 target, and consequently promotes apoptosis [34]. RP inhibits SCD mRNA expression through TP73 [35]. 17-β-estradiol induces SCD expression and the modulation of cellular lipid composition in ER+ cell lines and is necessary for estrogen-induced cell proliferation [36]. Overall, as these results showed, an anti-proliferative effect was provoked by RP treatment. Finally, RP also decreases mTOR-related protein levels [37–39]. Additionally, ACADSB, CCDC58, MPZL1 and SBSN protein expression predicts response to RP treatment. ACADSB affects valine and isoleucine metabolism [40], which is one of the pathways related to response to RP in flux activity analyses, as we will explain later. Probabilistic graphical models showed that RP treatment caused decreased node activity in mRNA processing. Additionally, metabolism A and B node activities accurately predict the response in cells treated with RP.

Proteomics coupled with gene ontology analyses allowed us to explore protein expression changes between

Table 3: Superoxide dismutase activity assay measurements

Cell line	Superoxide dismutase activity (%)
MCF7 Control	96.44%
MCF7 MTF	90.76%
CAMA1 Control	99.01%
CAMA1 MTF	97.09%
MDAMB231 Control	68.17%
MDAMB231 MTF	49.82%
HCC1143 Control	83.30%
HCC1143 MTF	86.44%

The experiment was performed in triplicate and one of the representative measurements is shown.

control and treated cells, suggesting that treatment with these drugs affects cell cycle progression. Therefore, the cell cycle was further assessed using flow cytometry. A cell cycle arrest in the G2/M phase was confirmed in all the MTF-treated cells except CAMA1, in which MTF had no effect on cell viability. Additionally, ER+ cells treated with RP (but not TNBC cells) had cell cycle arrest in G0/G1, which was confirmed at the cell proliferation level. It is known that mTOR controls cell cycle progression through S6K1 and 4E-BP1 [41]. Additionally, G0/G1 cell cycle arrest was previously described in MCF7 cells treated with RP [42]. Therefore, MTF and RP have cytostatic effects in breast cancer cell lines and cause a cell viability reduction, coupled with a disruption of the cell cycle. However, this response is diverse between various breast cancer cell lines.

On the other hand, FBA has traditionally been used in microbiology to study microorganism growth. This approach has recently been applied to study the Warburg effect [19]. We have developed a genome-scale cancer metabolic model that uses protein expression data to predict tumor growth rate. Previous studies have described cancer metabolic models using gene expression data [19, 43, 44]. Our model, however, used a whole human metabolism reconstruction and proteomics data to improve predictive accuracy. We assessed the model reliability by growth experimental studies in ER+ (MCF7 and T47D) and TNBC (MDAMB231 and MDAMB468) cells. This approach allows new hypotheses and provides a global vision of metabolism, and has been previously used to characterize metabolism in samples from patients with breast cancer, which enables us to address clinically relevant questions [10].

Model growth rate predictions were consistent with changes detected in viability assays in the cells treated with MTF. We explored the global flux for each pathway, calculating flux activities to identify metabolic pathways showing different behavior between the MTF-treated cells and the control cells. The pathways related to response to MTF treatment were glutamate and pyruvate metabolism. The pathways related to RP treatment response were valine, leucine and isoleucine metabolism and cholesterol metabolism. Although it is difficult to make comparisons between flux patterns, pathway flux activities could be a useful approach to understanding changes between various conditions.

Moreover, by using an FVA coupled with the Monte Carlo approach, an activation of enzymes related to ROS stress response associated with MTF treatment could be predicted. Catalase and SPODM activation by MTF have been described in other scenarios [45, 46] and, as previously mentioned, concurs in most cases with differences shown in protein expression, although this relationship is not always direct. For instance, SPODM showed a 1.25-fold increase in protein expression, but no increment at the flux level, because fluxes are conditioned not only by their own restrictions, but also by bounds from adjacent reactions. In addition, catalase and SPODM fluxes appear to be related to cell viability. For instance, CAMA1 cells treated with MTF did not show an increased catalase flux, perhaps due to the discrete effect of MTF treatment on CAMA1 viability. Some of these predictions have been verified in the SPODM activity assay. In general, SPODM activity measurements were consistent with FBA predictions. Variations between FBA predictions and SPODM activities could be due to the fact that FBA only take into account metabolic pathways. On the other hand, our model predicts an increase in nitric oxide synthase flux in MCF7 cells treated with MTF, as has been previously described in diabetic rats [47]. An increase in nitric oxide synthase implies a higher NO concentration, related to apoptosis processes and cytostatic effects in tumor cells, whereas low NO concentrations are associated with cell survival and proliferation [48]. This nitric oxide synthase activation could be related to the reduced proliferation observed in MCF7 cells treated with MTF. The fact that this effect was only predicted in MCF7 could be due to heterogeneity in the response mechanisms against this drug in various cellular contexts, and could be related to the observed differences in cell proliferation. It is remarkable that although no information about nitric oxide synthase abundance was provided by proteomics, our model reflects differences at the flux level in this

process, suggesting that both approaches, proteomics and flux balance analysis, offer complementary information.

To summarize our results, mitochondria and ATP binding node activities calculated by PGM functional nodes suggested that MTF effect takes place at mitochondria, a well-known fact [49]. As shown in FBA results, it also appears to increase ROS enzymes. Additionally, in MCF7 cells, an increase of nitric oxide synthase was predicted. Susceptibility to MTF treatment shown by MDAMB468 cells could be related to a *SLC22A1* SNP. As consequence of these events, MTF caused a heterogeneous effect on cell proliferation, consistent with a cell cycle arrest in the G2/M phase. .

On the other hand, RP treatment exerts greater effect on the cell proliferation of ER+ cells, mediated by a G0/G1 cell cycle arrest, as previously described [25]. This susceptibility of ER+ cell lines to RP treatment could be due to a SNP related to higher drug concentration. Finally, our results suggest that valine and isoleucine metabolism could be deregulated by RP treatment.

Our study has some limitations. FBA provides an optimal biomass value, but multiple combinations of fluxes leading to this optimum are possible, making assessing differential pathways between conditions difficult. In our study, this limitation was solved using resampling techniques; however, improvement of computational processes is still necessary. Regarding proteomics experiments, although they can improve model accuracy, because they allow direct measurement of enzyme levels, at this moment this approach can only provide values for about 57% of Recon2 reactions with the known GPR rule. Gene expression, however, with the limitation of being an indirect measurement of enzyme abundance, provides almost the full picture. Strikingly, FBA was not able to reflect cell viability changes due to RP treatment. Despite the potential of the FBA approach, it only takes into account differences at the metabolic level. It is well known that mTOR inhibition leads to massive changes in cell homeostasis; thus, it appears reasonable that modeling changes at the metabolism level alone could not predict these differences.

In this study, we propose a workflow to study response against drugs targeting metabolism using different experimental and computational methods that allow proposing new hypotheses and characterizing this response at molecular, functional and metabolic levels providing a whole vision of the process. Moreover, we have characterized differential protein expression patterns between cells treated with drugs targeting metabolism and control cells. We have also developed a computational workflow to evaluate the impact of metabolic alterations in tumor and cell growth rates, using proteomics data. Growth rates predicted by our model matched the viability results observed *in vitro* with drug exposure. In addition, probabilistic graphical models are useful to study effects related to biological processes instead of considering individual protein or gene expression patterns. Our holistic approach shows that various analyses provide complementary information, which can be used to suggest hypotheses about drug mechanisms of action and response that deserve subsequent validation. Finally, this type of analysis, when fully developed and validated, could be used to study metabolic patterns from tumor samples with a different response against drugs targeting metabolism.

MATERIALS AND METHODS

Cell culture and reagents

The ER+ breast cancer cell lines MCF7, T47D and CAMA1 and the triple-negative breast cancer cell lines MDAMB231, MDAMB468 and HCC1143 were cultured in RPMI-1640 medium with phenol red (Biological Industries), supplemented with 10% heat-inactivated fetal bovine serum (Gibco), 100 mg/mL penicillin (Gibco) and 100 mg/mL streptomycin (Gibco). All the cell lines were cultured at 37° C in a humidified atmosphere with 5% (v/v) CO_2 in the air. The MCF7, T47D and MDA-MB-231 cell lines were kindly provided by Dr. Nuria Vilaboa (La Paz University Hospital, previously obtained from ATCC in January 2014). The MDAMB468, CAMA1 and HCC1143 cell lines were obtained from ATCC (July 2014). Cell lines were routinely monitored in our laboratory and authenticated by morphology and growth characteristics, tested for Mycoplasma and frozen, and passaged for fewer than 6 months before experiments. The MTF (Sigma Aldrich D150959) and RP (Sigma Aldrich R8781) were obtained from Sigma-Aldrich (St. Louis, MO, USA).

Cell viability assays

The cells were treated with MTF and RP at a range of concentrations to establish an IC_{50} for each cell line. Approximately 5000 cells per well were seeded in 96-well plates. After 24 h, an appropriate concentration of drug was added to the cells, which were incubated for a total of 72 h. Untreated cells were used as a control. The CellTiter 96 AQueous One Solution Cell Proliferation Assay (Promega) kit was used for the quantification of cell survival after exposure to the drugs. After 72 h of incubation with the drug, CellTiter 96 AQueous One Solution was added to each well following the manufacturer's instructions, and absorbance was measured on a microplate reader (TECAN). Experiments were performed in triplicate. IC_{50} values were calculated using the Chou-Talalay method [50].

DNA extraction and SNP genotyping

DNA was extracted from untreated cells using the ISOLATE II RNA/DNA/Protein Kit (BIOLINE)

following manufacturer's instructions. We used TaqMan OpenArray technology on a QuantStudio 12K Flex Real-Time PCR System (Applied Biosystems®) with a custom SNP array format, which allows simultaneous genotyping of 180 SNPs in major drug metabolizing enzymes and transporters (PharmArray®). Information about the pharmacogenetic variants associated with RP and MTF response was gathered mostly from the variant and clinical annotations in the Pharmacogenomics Knowledge Base (PharmGKB; www.pharmgkb.org). The final selection of SNPs for our study was as follows: rs2032582, rs1045642, rs3213619 and rs1128503 in the *ABCB1* gene; rs55785340, rs4646438 and rs2740574 in *CYP3A4*; rs776746, rs55965422, rs10264272, rs41303343 and rs41279854 in *CYP3A5*; rs1057868 and rs2868177 in *POR* for RP; and rs55918055, rs36103319, rs34059508, rs628031, rs4646277, rs2282143, rs4646278, rs12208357 in *SLC22A1* and rs316019, rs8177516, rs8177517, rs8177507 and rs8177504 in *SLC22A2* for MTF. Molecular analyses for rs34130495 and rs2740574 were performed by classic sequencing because these probes were not originally included in our custom SNP array design.

Perturbation experiments

Suboptimal concentrations (IC_{70} or higher) were chosen in order to perform perturbation experiments (MTF 40 mM except for MDAMB468 20 m, RP 625 nM). Experiments were done per duplicate for each condition. Approximately 500,000 cells per well were seeded in 6-well plates. Twenty-four hours later, drugs against metabolism were added. After additional 24 h, proteins were extracted using the ISOLATE II RNA/DNA/Protein Kit (BIOLINE). Protein concentration was determined using the MicroBCA Protein Assay Kit (Pierce-Thermo Scientific). Protein extracts (10 µg) were digested with trypsin (Promega) (1:50). Peptides were desalted using in-house-produced C18 stage tips, then dried and resolubilized in 15 µl of 3% acetonitrile and 0.1% formic acid for MS analysis.

Liquid chromatography - mass spectrometry shotgun analysis

Mass spectrometry analysis was performed on a Q Exactive mass spectrometer coupled to a nano EasyLC 1000 (Thermo Fisher Scientific). Solvent composition at the two channels was 0.1% formic acid for channel A; and 0.1% formic acid, 99.9% acetonitrile for channel B. For each sample, 3 µL of peptides were loaded on a self-made column (75 µm × 150 mm) packed with reverse-phase C18 material (ReproSil-Pur 120 C18-AQ, 1.9 µm, Dr. Maisch GmbH) and eluted at a flow rate of 300 nL/min at a gradient from 2% to 35% B in 80 min, 47% B in 4 min and 98% B in 4 min. Samples were acquired in a randomized order. The mass spectrometer was operated in data-dependent mode, acquiring a full-scan MS spectra (300−1700 m/z) at a resolution of 70,000 at 200 m/z after accumulation to a target value of 3,000,000, followed by higher-energy collisional dissociation (HCD) fragmentation on the 12 most intense signals per cycle. The HCD spectra were acquired at a resolution of 35,000 using normalized collision energy of 25 and a maximum injection time of 120 ms. The automatic gain control was set to 50,000 ions. Charge state screening was enabled, and single and unassigned charge states were rejected. Only precursors with intensity above 8300 were selected for MS/MS (2% underfill ratio). Precursor masses previously selected for MS/MS measurement were excluded from further selection for 30 s, and the exclusion window was set at 10 ppm. The samples were acquired using internal lock mass calibration on m/z 371.1010 and 445.1200.

Protein identification and label-free protein quantification

The acquired raw MS data were processed by MaxQuant (version 1.4.1.2), followed by protein identification using the integrated Andromeda search engine. Each file is kept separate in the experimental design to obtain individual quantitative values. The spectra were searched against a forward Swiss-Prot human database, concatenated to a reversed decoyed FASTA database and common protein contaminants (NCBI taxonomy ID9606, release date 2014-05-06). Methionine oxidation and N-terminal protein acetylation were set as variable modification. Enzyme specificity was set to trypsin/P allowing a minimal peptide length of 7 amino acids and a maximum of two missed cleavages. Precursor and fragment tolerance was set to 10 ppm and 20 ppm, respectively, for the initial search. The maximum false discovery rate (FDR) was set to 0.01 for peptides and 0.05 for proteins. Label-free quantification was enabled, and a 2-minute window for match between runs was applied. The requantify option was selected. For protein abundance, the intensity (Intensity) as expressed in the protein groups file was used, corresponding to the sum of the precursor intensities of all identified peptides for the respective protein group. Only quantifiable proteins (defined as protein groups showing two or more razor peptides) were considered for subsequent analyses. Protein expression data were transformed (hyperbolic arcsine transformation), and missing values (zeros) were imputed using the missForest R package [51]. The protein intensities were normalized by scaling the median protein intensity in each sample to the same values. Then values were $\log_2$ transformed.

All the mass spectrometry raw data files acquired in this study may be downloaded from Chorus (http://chorusproject.org) under the project name "Metabolism targeting in breast cancer cells". The peptides output file

from the MaxQuant analysis is provided as supplementary material (Supplementary Table 2).

Gene ontology analyses

Protein expression patterns were compared between the control and treated cells, and deltas were calculated for each drug in each cell line by subtracting control protein expression from treated cell protein expression values. Gene ontology analyses were performed to determine differential functions between the control and the treated cells. For this, we selected protein showing a change in expression values (delta) higher than 1.5 or lower than -1.5; this delta value was calculated for each protein as the treated cell expression value minus the control cell expression value. Protein-to-gene ID conversion were performed using Uniprot (http://www.uniprot.org) and DAVID [52]. The gene ontology analyses were performed using the functional annotation chart tool provided by DAVID. We used "homo sapiens" as a background list and selected only GOTERM-FAT gene ontology categories and Biocarta and KEGG pathways. Functional categories with $p < .05$ and a FDR below 5% were considered as significant.

Probabilistic graphical models, functional node activity measurements and response predicted models

Network construction was performed using probabilistic graphical models compatible with high dimensional data using correlation coefficients as associative measures as previously described [9]. To build this model, protein expression data without other *a priori* information was used. *grapHD* package [53] and R v3.2.5 [54] were employed to build the model.

The resulting network was split into several branches and a gene ontology analysis was used to explore the major biological function for each branch, defining functional nodes. Again, gene ontology analyses were performed in DAVID webtool [52] using "homo sapiens" as background and GOTERM-FAT, Biocarta and KEGG categories. Functional node activity was calculated as the mean delta between treated and untreated cells of all proteins related to the assigned majority node function. In order to relate drug response to functional processes, multiple linear regression models were performed using IBM SPSS Statistics.

Cytometry experiments

Some 500,000 cells were seeded in each well per duplicate. Twenty-four hours later, drugs were added and, after 72 h, the cells were fixed in ethanol and marked with propidium iodide. Cells were acquired using a FACScan cytometer equipped with a blue laser at a wavelength of 488 nm. Acquired data were analyzed using BD CellQuest Pro software, first filtering cells by size and complexity in order to exclude debris, and then excluding doublets and triplets by FL2-W/FL2-A.

Flux balance analysis and E-flux algorithm

FBA was used to build a metabolic computational model that predicts growth rates. FBA calculates the flow of metabolites through metabolic networks and predicts growth rates or the rate of production of a given metabolite. It was performed using the COBRA Toolbox v2.0 [55] available for MATLAB and the human metabolism reconstruction Recon2 [56]. MATLAB R2014b and *glpk* solver were used. The biomass reaction proposed in Recon2 was used as an objective function representative of growth rate in tumor cells. Proteomics expression data were included in the model by solving GPR rules and using a modified E-flux algorithm [57]. Measuring GPR rule estimation values was performed using a variation of the method described by Barker *et al*. [58]. As described in previous works [10], the mathematical operations used to calculate the numerical value were the sum of "OR" expressions and the minimum of "AND" expressions. Finally, the GPR rule values, *aj*, were normalized to a [0, 1] interval, using a uniform distribution formula. The normalized values have been used to establish both new lower and upper reaction bounds. If the reaction is irreversible the new bounds are 0 and *aj*, and if the reaction is reversible the new bounds are - *aj* and *aj* (Supplementary Figure 6, Supplementary File 13).

Metabolism model validation

In order to validate model predictions we used dynamic FBA, which allows the prediction of cell growth during a period of time [43] and experimental growth studies of cell lines were performed. Dynamic FBA consists of an iterative approach based on a quasi-steady state assumption [59]. MCF7, T47D, MDAMB468 and MDAMB231 were seeded at an initial cell density of 1,000,000 cells. Cells within the same area were counted once a day for 3 days. To perform the dynamic FBA, experimental cell density at the beginning and experimental measured glucose concentration in the medium were used as inputs in the computational simulation. Glucose presented in the medium was measured using an ABL90 FLEX blood analyzer (Radiometer). *dynamicFBA* function implemented in COBRA Toolbox was used. The simulation was performed for a time of 72 hours as the cell density experimental measurements.

Flux activities

With the aim of comparing the activity of the various pathway fluxes between the control and the treated cells, flux activity was calculated for each condition. Flux activity was defined by the sum of fluxes for all reactions involved in one pathway as defined in the Recon2. Then, linear regression models were performed.

Flux variability analysis and the Monte Carlo approach

One obvious limitation to the FBA approach is that this analysis provides a unique optimal tumor growth rate, however, multiple combinations of fluxes can lead to this optimal value. In order to evaluate a representative sample of these multiple solutions, a Monte Carlo approach [60] was used to compare differential fluxes between treated and untreated cells. The solution showing the maximum sum of all the fluxes was then used to calculate the flux change between the control and the treated cells. This criterion was selected under the premise that if a protein was experimentally measured it was because that protein was going to be used by the cell; thus, maximum flux solution picks up all measured proteins. On the other hand, FVA provides the possible maximum and minimum fluxes for each reaction; therefore, the flux range for each reaction. This range was used to calculate the flux change between the control and the treated cells for a given reaction as a percentage of the flux range for that reaction. Reactions showing a flux change between the control and the treated cells over 95% of this range were identified for each condition. Monte Carlo results for these reactions were used to check if maximum solution flux is representative of the most frequent solution flux for this reaction.

Superoxide dismutase activity assay

To validate some of our model hypotheses, a SPODM activity assay was performed in triplicate, using the Superoxide Dismutase Assay Kit (Sigma-Aldrich, 19160). Some 500,000 cells per well were seeded, and after 24 h, MTF was added at 40 mM (except for the MDAMB468 cell line, in which a 20 mM concentration was used). Twenty-four hours later, SPODM activities were measured following the manufacturer's instructions.

Statistical analyses and software suites

Dose-response curves were constructed with GraphPad Prism 6. Gene and protein interactions for each drug were obtained from the Comparative Toxicogenomics Database (http://ctdbase.org/) [61]. Linear and multiple regression models were built using IBM SPSS Statistics.

Abbreviations

MTF: Metformin; RP: Rapamycin; ER+: Hormone-receptor positive; TNBC: Triple negative; SNP: Single Nucleotide Polymorphism; FBA: Flux Balance Analysis; ROS: reactive oxygen species; GPR: Gen-Protein-Reaction; FVA: Flux Variability Analysis; NO: nitric oxide; SPODM: superoxide dismutase.

Author contributions

All the authors have directly participated in the preparation of this manuscript and have approved the final version submitted. LT-F and RL-V performed the viability experiments and prepared the proteomics samples. LT-F performed the FBA. PN contributed the mass spectrometry data. JMA, HN and PM contributed the probabilistic graphical models. MD-A contributed the GPR rule calculation. LT-F, RL-V and RA-L contributed the enzyme activity assays. LT-F, GP-V, AZ-M and SL-A performed the statistical analysis, the probabilistic graphical model interpretation and the gene ontology analyses. ID, PA and AMB contributed the SNP study. LT-F drafted the manuscript. LT-F, AG-P, JAFV, JF and EE conceived of the study and participated in its design and interpretation. AG-P, JAFV, and EE supported the manuscript drafting. AG-P and JAFV coordinated the study. All the authors have read and approved the final manuscript.

ACKNOWLEDGMENTS

The cytometry experiments were performed at the Cytometry and Fluorescence Microscopy Center, Faculty of Chemistry, Complutense University of Madrid, Spain.

CONFLICTS OF INTEREST

JAFV, EE and AG-P are shareholders in Biomedica Molecular Medicine SL. LT-F is an employee of Biomedica Molecular Medicine SL. The other authors declare no competing interests.

FUNDING

This study was supported by Instituto de Salud Carlos III, Spanish Economy and Competitiveness Ministry, Spain and co-funded by the FEDER program, "Una forma de hacer Europa" (PI15/01310). LT-F is supported by the Spanish Economy and Competitiveness Ministry (DI-15-07614). The funders had no role in the study design, data collection and analysis, decision to publish or preparation of the manuscript.

REFERENCES

1. Siegel RL, Miller KD, Jemal A. Cancer statistics, 2015. CA Cancer J Clin. 2015; 65:5–29.
2. Fisher B, Costantino J, Redmond C, Poisson R, Bowman D, Couture J, Dimitrov NV, Wolmark N, Wickerham DL, Fisher ER, Margolese R, Robidoux A, Shibata H, et al. A randomized clinical trial evaluating tamoxifen in the treatment of patients with node-negative breast cancer who have estrogen-receptor-positive tumors. N Engl J Med. 1989; 320:479–84. https://doi.org/10.1056/nejm198902233200802.
3. Buzdar A, Jonat W, Howell A, Jones SE, Blomqvist C, Vogel CL, Eiermann W, Wolter JM, Azab M, Webster A, Plourde PV, and Arimidex Study Group. Anastrozole, a potent and selective aromatase inhibitor, versus megestrol acetate in postmenopausal women with advanced breast cancer: results of overview analysis of two phase III trials. J Clin Oncol. 1996; 14:2000–11.
4. Slamon DJ, Clark GM, Wong SG, Levin WJ, Ullrich A, McGuire WL. Human breast cancer: correlation of relapse and survival with amplification of the HER-2/neu oncogene. Science. 1987; 235:177–82.
5. Perou CM, Sorlie T, Eisen MB, van de Rijn M, Jeffrey SS, Rees CA, Pollack JR, Ross DT, Johnsen H, Akslen LA, Fluge O, Pergamenschikov A, Williams C, et al. Molecular portraits of human breast tumours. Nature. 2000; 406:747–52. https://doi.org/10.1038/35021093.
6. Hanahan D, Weinberg RA. Hallmarks of cancer: the next generation. Cell. 2011; 144:646–74. https://doi.org/10.1016/j.cell.2011.02.013.
7. Warburg O. The metabolism of carcinoma cells. J Cancer Res. 1925; 9:148–63.
8. DeBerardinis RJ, Mancuso A, Daikhin E, Nissim I, Yudkoff M, Wehrli S, Thompson CB. Beyond aerobic glycolysis: transformed cells can engage in glutamine metabolism that exceeds the requirement for protein and nucleotide synthesis. Proc Natl Acad Sci U S A. 2007; 104:19345–50. https://doi.org/10.1073/pnas.0709747104.
9. Gámez-Pozo A, Berges-Soria J, Arevalillo JM, Nanni P, López-Vacas R, Navarro H, Grossmann J, Castaneda CA, Main P, Díaz-Almirón M, Espinosa E, Ciruelos E, Fresno Vara JA. Combined label-free quantitative proteomics and microRNA expression analysis of breast cancer unravel molecular differences with clinical implications. Cancer Res. 2015; 75:2243–53.
10. Gámez-Pozo A, Trilla-Fuertes L, Berges-Soria J, Selevsek N, López-Vacas R, Díaz-Almirón M, Nanni P, Arevalillo JM, Navarro H, Grossmann J, Gayá Moreno F, Gómez Rioja R, Prado-Vázquez G, et al. Functional proteomics outlines the complexity of breast cancer molecular subtypes. Sci Rep. 2017; 7:10100. https://doi.org/10.1038/s41598-017-10493-w.
11. Jones NP, Schulze A. Targeting cancer metabolism--aiming at a tumour's sweet-spot. Drug Discov Today. 2012; 17:232–41. https://doi.org/10.1016/j.drudis.2011.12.017.
12. Zhou G, Myers R, Li Y, Chen Y, Shen X, Fenyk-Melody J, Wu M, Ventre J, Doebber T, Fujii N, Musi N, Hirshman MF, Goodyear LJ, et al. Role of AMP-activated protein kinase in mechanism of metformin action. J Clin Invest. 2001; 108:1167–74. https://doi.org/10.1172/JCI13505.
13. Beck JT. Potential role for mammalian target of rapamycin inhibitors as first-line therapy in hormone receptor-positive advanced breast cancer. Onco Targets Ther. 2015; 8:3629–38. https://doi.org/10.2147/OTT.S88037.
14. Varma A, Palsson BO. Parametric sensitivity of stoichiometric flux balance models applied to wild-type Escherichia coli metabolism. Biotechnol Bioeng. 1995; 45:69–79. https://doi.org/10.1002/bit.260450110.
15. Pramanik J, Keasling JD. Stoichiometric model of Escherichia coli metabolism: incorporation of growth-rate dependent biomass composition and mechanistic energy requirements. Biotechnol Bioeng. 1997; 56:398–421. https://doi.org/10.1002/(SICI)1097-0290(19971120)56:4<398::AID-BIT6>3.0.CO;2-J.
16. Edwards J. Functional genomics and the computational analysis of bacterial metabolism. Department of Bioengineering. San Diego: University of California. 1999.
17. Edwards JS, Ibarra RU, Palsson BO. In silico predictions of Escherichia coli metabolic capabilities are consistent with experimental data. Nat Biotechnol. 2001; 19:125–30.
18. Schilling C, Palsson B. The underlying pathway structure of biochemical reaction networks. Proc Natl Acad Sci USA). 1998; 4193–8.
19. Asgari Y, Zabihinpour Z, Salehzadeh-Yazdi A, Schreiber F, Masoudi-Nejad A. Alterations in cancer cell metabolism: the Warburg effect and metabolic adaptation. Genomics. 2015; 105:275–81.
20. Wortham M, Czerwinski M, He L, Parkinson A, Wan YJ. Expression of constitutive androstane receptor, hepatic nuclear factor 4 alpha, and P450 oxidoreductase genes determines interindividual variability in basal expression and activity of a broad scope of xenobiotic metabolism genes in the human liver. Drug Metab Dispos. 2007; 35:1700–10. https://doi.org/10.1124/dmd.107.016436.
21. Motamedian E, Taheri E, Bagheri F. Proliferation inhibition of cisplatin-resistant ovarian cancer cells using drugs screened by integrating a metabolic model and transcriptomic data. Cell Prolif. 2017; 50:e12370. https://doi.org/10.1111/cpr.12370.
22. Buler M, Aatsinki SM, Izzi V, Uusimaa J, Hakkola J. SIRT5 is under the control of PGC-1α and AMPK and is involved in regulation of mitochondrial energy metabolism. FASEB J. 2014; 28:3225–37. https://doi.org/10.1096/fj.13-245241.
23. Gomes P, Outeiro TF, Cavadas C. Emerging Role of Sirtuin 2 in the Regulation of Mammalian Metabolism. Trends

Pharmacol Sci. 2015; 36:756–68. https://doi.org/10.1016/j.tips.2015.08.001.

24. Lin ZF, Xu HB, Wang JY, Lin Q, Ruan Z, Liu FB, Jin W, Huang HH, Chen X. SIRT5 desuccinylates and activates SOD1 to eliminate ROS. Biochem Biophys Res Commun. 2013; 441:191–5. https://doi.org/10.1016/j.bbrc.2013.10.033.

25. Cuyàs E, Corominas-Faja B, Joven J, Menendez JA. Cell cycle regulation by the nutrient-sensing mammalian target of rapamycin (mTOR) pathway. Methods Mol Biol. 2014; 1170:113–44. https://doi.org/10.1007/978-1-4939-0888-2_7.

26. Tang LH, Contractor T, Clausen R, Klimstra DS, Du YC, Allen PJ, Brennan MF, Levine AJ, Harris CR. Attenuation of the retinoblastoma pathway in pancreatic neuroendocrine tumors due to increased cdk4/cdk6. Clin Cancer Res. 2012; 18:4612–20. https://doi.org/10.1158/1078-0432.CCR-11-3264.

27. Knudsen ES, Witkiewicz AK. Defining the transcriptional and biological response to CDK4/6 inhibition in relation to ER+/HER2- breast cancer. Oncotarget. 2016; 7:69111–23. https://doi.org/10.18632/oncotarget.11588.

28. Gonzalez J, Harris T, Childs G, Prystowsky MB. Rapamycin blocks IL-2-driven T cell cycle progression while preserving T cell survival. Blood Cells Mol Dis. 2001; 27:572–85. https://doi.org/10.1006/bcmd.2001.0420.

29. Wang XC, Tian J, Tian LL, Wu HL, Meng AM, Ma TH, Xiao J, Xiao XL, Li CH. Role of Cks1 amplification and overexpression in breast cancer. Biochem Biophys Res Commun. 2009; 379:1107–13. https://doi.org/10.1016/j.bbrc.2009.01.028.

30. Cui Y, Huang Q, Auman JT, Knight B, Jin X, Blanchard KT, Chou J, Jayadev S, Paules RS. Genomic-derived markers for early detection of calcineurin inhibitor immunosuppressant-mediated nephrotoxicity. Toxicol Sci. 2011; 124:23–34. https://doi.org/10.1093/toxsci/kfr217.

31. Pannu V, Rida PC, Ogden A, Turaga RC, Donthamsetty S, Bowen NJ, Rudd K, Gupta MV, Reid MD, Cantuaria G, Walczak CE, Aneja R. HSET overexpression fuels tumor progression via centrosome clustering-independent mechanisms in breast cancer patients. Oncotarget. 2015; 6:6076–91. https://doi.org/10.18632/oncotarget.3475.

32. Davies TH, Ning YM, Sánchez ER. Differential control of glucocorticoid receptor hormone-binding function by tetratricopeptide repeat (TPR) proteins and the immunosuppressive ligand FK506. Biochemistry. 2005; 44:2030–8. https://doi.org/10.1021/bi048503v.

33. Vilasco M, Communal L, Mourra N, Courtin A, Forgez P, Gompel A. Glucocorticoid receptor and breast cancer. Breast Cancer Res Treat. 2011; 130:1–10. https://doi.org/10.1007/s10549-011-1689-6.

34. He H, Sun Y. Ribosomal protein S27L is a direct p53 target that regulates apoptosis. Oncogene. 2007; 26:2707–16. https://doi.org/10.1038/sj.onc.1210073.

35. Rosenbluth JM, Mays DJ, Jiang A, Shyr Y, Pietenpol JA. Differential regulation of the p73 cistrome by mammalian target of rapamycin reveals transcriptional programs of mesenchymal differentiation and tumorigenesis. Proc Natl Acad Sci USA. 2011; 108:2076–81. https://doi.org/10.1073/pnas.1011936108.

36. Belkaid A, Duguay SR, Ouellette RJ, Surette ME. 17β-estradiol induces stearoyl-CoA desaturase-1 expression in estrogen receptor-positive breast cancer cells. BMC Cancer. 2015; 15:440. https://doi.org/10.1186/s12885-015-1452-1.

37. Boulay A, Rudloff J, Ye J, Zumstein-Mecker S, O'Reilly T, Evans DB, Chen S, Lane HA. Dual inhibition of mTOR and estrogen receptor signaling *in vitro* induces cell death in models of breast cancer. Clin Cancer Res. 2005; 11:5319–28. https://doi.org/10.1158/1078-0432.CCR-04-2402.

38. O'Reilly T, McSheehy PM, Wartmann M, Lassota P, Brandt R, Lane HA. Evaluation of the mTOR inhibitor, everolimus, in combination with cytotoxic antitumor agents using human tumor models *in vitro* and *in vivo*. Anticancer Drugs. 2011; 22:58–78. https://doi.org/10.1097/CAD.0b013e3283400a20.

39. Yee KW, Zeng Z, Konopleva M, Verstovsek S, Ravandi F, Ferrajoli A, Thomas D, Wierda W, Apostolidou E, Albitar M, O'Brien S, Andreeff M, Giles FJ. Phase I/II study of the mammalian target of rapamycin inhibitor everolimus (RAD001) in patients with relapsed or refractory hematologic malignancies. Clin Cancer Res. 2006; 12:5165–73. https://doi.org/10.1158/1078-0432.CCR-06-0764.

40. Andresen BS, Christensen E, Corydon TJ, Bross P, Pilgaard B, Wanders RJ, Ruiter JP, Simonsen H, Winter V, Knudsen I, Schroeder LD, Gregersen N, Skovby F. Isolated 2-methylbutyrylglycinuria caused by short/branched-chain acyl-CoA dehydrogenase deficiency: identification of a new enzyme defect, resolution of its molecular basis, and evidence for distinct acyl-CoA dehydrogenases in isoleucine and valine metabolism. Am J Hum Genet. 2000; 67:1095–103. https://doi.org/10.1086/303105.

41. Fingar DC, Richardson CJ, Tee AR, Cheatham L, Tsou C, Blenis J. mTOR controls cell cycle progression through its cell growth effectors S6K1 and 4E-BP1/eukaryotic translation initiation factor 4E. Mol Cell Biol. 2004; 24:200–16.

42. Tengku Din TA, Seeni A, Khairi WN, Shamsuddin S, Jaafar H. Effects of rapamycin on cell apoptosis in MCF-7 human breast cancer cells. Asian Pac J Cancer Prev. 2014; 15:10659–63.

43. Resendis-Antonio O, Checa A, Encarnación S. Modeling core metabolism in cancer cells: Surveying the topology underling the Warburg effect. PLoS One. 2010; 5:e12383.

44. Vázquez A, Liu J, Zhou Y, Oltvai ZN. Catabolic efficiency of aerobic glycolysis: the Warburg effect revisited. BMC Syst Biol. 2010; 4:58. https://doi.org/10.1186/1752-0509-4-58.

45. Dai J, Liu M, Ai Q, Lin L, Wu K, Deng X, Jing Y, Jia M, Wan J, Zhang L. Involvement of catalase in the protective benefits of metformin in mice with oxidative liver injury. Chem Biol Interact. 2014; 216:34–42. https://doi.org/10.1016/j.cbi.2014.03.013.

46. Kukidome D, Nishikawa T, Sonoda K, Imoto K, Fujisawa K, Yano M, Motoshima H, Taguchi T, Matsumura T, Araki E. Activation of AMP-activated protein kinase reduces hyperglycemia-induced mitochondrial reactive oxygen species production and promotes mitochondrial biogenesis in human umbilical vein endothelial cells. Diabetes. 2006; 55:120–7.

47. Volarevic V, Misirkic M, Vucicevic L, Paunovic V, Simovic Markovic B, Stojanovic M, Milovanovic M, Jakovljevic V, Micic D, Arsenijevic N, Trajkovic V, Lukic ML. Metformin aggravates immune-mediated liver injury in mice. Arch Toxicol. 2015; 89:437–50. https://doi.org/10.1007/s00204-014-1263-1.

48. Vannini F, Kashfi K, Nath N. The dual role of iNOS in cancer. Redox Biol. 2015; 6:334–43. https://doi.org/10.1016/j.redox.2015.08.009.

49. El-Mir MY, Nogueira V, Fontaine E, Avéret N, Rigoulet M, Leverve X. Dimethylbiguanide inhibits cell respiration via an indirect effect targeted on the respiratory chain complex I. J Biol Chem. 2000; 275:223–8.

50. Chou TC. Theoretical basis, experimental design, and computerized simulation of synergism and antagonism in drug combination studies. Pharmacol Rev. 2006; 58:621–81. https://doi.org/10.1124/pr.58.3.10.

51. Stekhoven DJ, Bühlmann P. MissForest—nonparametric missing value imputation for mixed-type data. Bioinformatics. 2012; 28:112–8. https://doi.org/10.1093/bioinformatics/btr597.

52. Huang W, Sherman BT, Lempicki RA. Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. Nat Protoc. 2009; 4:44–57. https://doi.org/10.1038/nprot.2008.211.

53. Abreu G, Edwards D, Labouriau R. High-Dimensional Graphical Model Search with the gRapHD R Package. J Stat Softw. 2010; 37: 1–18.

54. R Core Team. R: A language and environment for statistical computing. Vienna, Austria: R Foundation for Statistical Computing. 2013.

55. Schellenberger J, Que R, Fleming RM, Thiele I, Orth JD, Feist AM, Zielinski DC, Bordbar A, Lewis NE, Rahmanian S, Kang J, Hyduke DR, Palsson BØ. Quantitative prediction of cellular metabolism with constraint-based models: the COBRA Toolbox v2.0. Nat Protoc. 2011; 6:1290–307.

56. Thiele I, Swainston N, Fleming RM, Hoppe A, Sahoo S, Aurich MK, Haraldsdottir H, Mo ML, Rolfsson O, Stobbe MD, Thorleifsson SG, Agren R, Bölling C, et al. A community-driven global reconstruction of human metabolism. Nat Biotechnol. 2013; 31:419–25. https://doi.org/10.1038/nbt.2488.

57. Colijn C, Brandes A, Zucker J, Lun D, Weiner B, Farhat M, Cheng T, Moody B, Murray M, Galagan J. Interpreting expression data with metabolic flux models: Predicting Mycobacterium tuberculosis mycolic acid production. PLoS Comput Biol. 2009; 5:e1000489.

58. Barker BE, Sadagopan N, Wang Y, Smallbone K, Myers CR, Xi H, Locasale JW, Gu Z. A robust and efficient method for estimating enzyme complex abundance and metabolic flux from expression data. Comput Biol Chem. 2015; 59:98–112. https://doi.org/10.1016/j.compbiolchem.2015.08.002.

59. Varma A, Palsson BO. Stoichiometric flux balance models quantitatively predict growth and metabolic by-product secretion in wild-type Escherichia coli W3110. Appl Environ Microbiol. 1994; 60:3724–31.

60. Schellenberger J. Monte Carlo simulation in Systems Biology. Bioinformatics and Systems Biology. 2010: 162.

61. Davis AP, Grondin CJ, Johnson RJ, Sciaky D, King BL, McMorran R, Wiegers J, Wiegers TC, Mattingly CJ. The Comparative Toxicogenomics Database: update 2017. Nucleic Acids Res. 2017; 45:D972–D8. https://doi.org/10.1093/nar/gkw838.

Computational metabolism modeling predicts risk of distant relapse-free survival in breast cancer patients

Lucía Trilla-Fuertes[1¶], Angelo Gámez-Pozo[1,2¶], Mariana Díaz- Almirón[3], Guillermo Prado-Vázquez[1], Andrea Zapater-Moros[2], Rocío López-Vacas[2], Paolo Nanni[4], Pilar Zamora[5], Enrique Espinosa[5], and Juan Ángel Fresno Vara[2*].

[1] Biomedica Molecular Medicine SL, C/ Faraday 7, 28049, Madrid, Spain

[2] Molecular Oncology & Pathology Lab, Institute of Medical and Molecular Genetics-INGEMM, La Paz University Hospital-IdiPAZ, Paseo de la Castellana 261, 28046, Madrid, Spain

[3] Biostatistics Unit, La Paz University Hospital-IdiPAZ, Paseo de la Castellana 261, 28046, Madrid. Spain

[4] Functional Genomics Centre Zurich, University of Zurich/ETH Zurich, Winterthurerstrasse 190, 8057, Zurich, Switzerland.

[5] Medical Oncology Service, La Paz University Hospital-IdiPAZ, Paseo de la Castellana 261, 28046, Madrid, Spain

[¶] These authors contribute equally to this work

* Corresponding author

e-mails: LT-F: lucia.trilla@biomedicamm.com; AG-P: angelo.gamez@idipaz.es; MD-A: manadiaz2@gmail.com; GP-V: 14gprado@gmail.com; AZ-M: andreazapater@hotmail.com; RL-V: rlvacas@gmail.com; PN: paolo.nanni@fgcz.uzh.ch; PZ: zamorapilar@gmail.com; EE: eespinosa00@hotmail.com; JAFV: juanangel.fresno@salud.madrid.org

Abstract

Aims: Differences in metabolism among breast cancer subtypes suggest that metabolism plays an important role in this disease. Flux Balance Analysis is used to explore these differences as well as drug response.

Materials & Methods: Proteomics data from breast tumors were obtained by mass-spectrometry. Flux Balance Analysis was performed to study metabolic networks. Flux activities from metabolic pathways were calculated and used to build prognostic models.

Results: Flux activities of vitamin A, tetrahydrobiopterin and beta-alanine metabolism pathways split our population into low- and high-risk patients. Additionally, flux activities of glycolysis and glutamate metabolism split triple negative tumors into low- and high-risk groups.

Conclusions: Flux activities summarize Flux Balance Analysis data and can be associated with prognosis in cancer.

Keywords: breast cancer, metabolism, prognosis, flux balance analysis, personalized medicine

Introduction

Breast cancer has a high incidence, with 266,120 estimated new cases and 40,920 estimated deaths in women in the United States during 2018 [1]. The expression of estrogen receptor, progesterone receptor and human epidermal growth factor receptor 2 (HER2) classifies breast cancer into one of three groups: hormone receptor positive / HER2 negative (ER+), HER2 positive (HER2+) or triple negative (TNBC). In previous works, we defined a new subtype within ER+ tumors with a clinical outcome and molecular features more similar to TNBC. This new subtype was called TN-like. ER+ tumors which still have ER+ characteristics were renamed as ER-true [2].

Reprogramming of metabolism is a hallmark of cancer [3]. Tumor cells use glucose to produce lactate, thus avoiding glucose metabolism through the Krebs cycle [4]. Tumor cells also produce lactate from glutamine and then generate NADPH in a process known as glutaminolysis [5]. However, not all tumors show the same metabolic alterations. In previous studies we described that there are differences in metabolism between breast cancer subtypes [6].

Proteomics provides detailed information about biological processes. Technical improvements currently allow the quantification of thousands of proteins. This information can be used to calculate the output of metabolic pathways with Flux Balance Analysis (FBA). FBA is a computational method used to study metabolic networks, making possible to predict tumor growth rate or the rate of production of a metabolite [7]. In our previous study, preliminary results of FBA were shown. FBA predicted a higher growth rate in TN-like tumors than in ER-true. Moreover, tumor growth predictions for TNBC and TN-like were comparable. We have proposed the calculation of flux activities as a summary measurement of flux distributions and showed that flux activities can be used to compare metabolic patterns between tumors or cells [8].

The aim of this work is to study in depth the metabolic differences previously characterized in breast cancer [2,6]. More specifically, proteomics data were analyzed through FBA to find metabolic pathways with prognostic value.

Materials and Methods

Patient cohort

One hundred and six formalin-fixed paraffin-embedded (FFPE) samples from I+12 Biobank and from IdiPAZ Biobank, both integrated in the Spanish Hospital Biobank Network (RetBioH; www.redbiobancos.es), were analyzed. This study was approved by the Ethical Committees of Hospital 12 de Octubre and Hospital La Paz.

Samples were selected according to these criteria: 1) node-positive disease, 2) no HER2 overexpression, and 3) patients treated with adjuvant chemotherapy and hormonal therapy in the case of ER+ tumors. The following clinical data were recorded: patient's age, tumor size, lymph node status, tumor grade, adjuvant therapy administered and distant metastasis-free survival.

Protein isolation

Proteins were isolated from FFPE samples as described in previous works [6,9]. Briefly, FFPE slices were deparaffinized in xylene and washed with absolute ethanol. Extracts were prepared in 2% SDS buffer using a protocol based on heat-induced antigen retrieval [10]. Protein concentration was determined using the MicroBCA Protein Assay Kit (Pierce-Thermo Scientific). Then, protein extracts were digested by trypsin and SDS was removed. Samples were dried and resolubilized in 15 µL of a 0.1% formic acid and 3% acetonitrile solution.

Mass-spectrometry experiments and protein identification

Samples were analyzed on a LTQ-Orbitrap Velos hybrid mass spectrometer (Thermo Fischer Scientific, Bremen, Germany) coupled to NanoLC-Ultra system (Eksigent Technologies, Dublin, CA, USA). Peptides were separated on a self-made column. Solvent composition was 0.1% formic acid for channel A, and 0.1% formic acid and 99.9% acetonitrile for channel B. The mass-spectrometer was in data-dependent mode, acquiring full-scan spectra at a resolution of 30, 000 at 400 m/z after accumulation to a target value of 1,000,000, followed by CID (collision-induced dissociation) fragmentation on the twenty most intense signals per cycle.

The acquired data was analyzed by MaxQuant (version 1.2.7.4), and protein identification was done using Andromeda. Oxidation (M), deamidation (N, Q) and N-terminal protein acetylation was set as modifications. Protein abundance was calculated based on the normalized spectra intensity (LFQ).

Proteomics data were transformed into log2 and missing values were imputed to a normal distribution using Perseus [11]. Additionally, quality criteria as presence in at least 75% of the samples of each type or at least two unique peptides were required.

Proteomics data is publicly available in Chorus repository (http://chorusproject.org) under the name "Breast Cancer Proteomics".

Flux Balance Analysis and flux activities

FBA was performed using COBRA Toolbox available for MATLAB and the whole human metabolic reconstruction Recon2 [12,13]. The biomass reaction proposed in the Recon2 was used as an objective function to quantify tumor growth rate. Gene-Protein-Reaction rules (GPR), which relate genes and proteins with the enzymes involved in the reactions contained into the Recon2, were solved using a modification of the algorithm by Barker et al. [14]. Briefly, "OR" expressions are treated as a sum and "AND" expressions are calculated as the minimum [2,8]. Then, GPR values were introduced into the model using a modified E-flux consisted on normalize GPR data using a normal distribution formula [15].

Flux activities were calculated as in previous works [8]. Briefly, the flux activities were the sum of fluxes for all the reactions included in a metabolic pathway defined in the Recon2.

Statistical analyses

Distant metastasis-free survival was selected as the prognostic variable. Statistical analyses were performed using GraphPad Prism v6 and predictors for distant metastasis-free survival was built using BRB Array Tool, developed by Dr. Richard Simon's team. Cox regression models were done in SPSS IBM Statistics v20.

Results

Patient characteristics

One hundred and six samples were selected for the study. Finally, ninety-six patients diagnosed of breast cancer were analyzed. This cohort had been used in previous works from our group [2,6].

After performing the analyses of the 96 samples, of the 71 ER+ samples, 50 were reclassified as ER-true and 21 as TN-like, and twenty-six were TNBC tumors [2].

Proteomics experiments

One hundred and six patients were recruited in this study. Four samples did not have enough protein amounts to perform mass-spectrometry (MS) experiments. After MS experiments, ninety-six samples had useful protein expression data. 3,239 proteins were measured. Quality criteria and filters were applied and a total of 1,095 proteins had two unique peptides and were detected in at least 75% of the samples of one type, i.e. ER+ or TNBC.

Flux activities

FBA is a computational method used to study metabolic networks. The whole metabolic human reconstruction Recon2 was used to perform these analyses. The Recon2 is composed by 7,440 reactions and 5,063 metabolites grouped in 101 different metabolic pathways.

Using the results from FBA, flux activities were calculated as the sum of fluxes for the reactions included in each metabolic pathway described in the Recon2. Comparing breast cancer subtypes, significant differences in nucleotide interconversion pathway and reactive-oxygen species (ROS) detoxification were found between ER-true and TNBC and between ER-true and TN-like respectively (Figure 1).

Prognostic signatures using flux activities

In order to identify flux activities related with distant relapse-free survival, BRB Array Tool from Dr. Richard Simon was used. We found five flux activities related with distant relapse-free survival (Sup Table 1).

Using these flux activities, a predictor of distant metastasis-free survival was constructed. This signature included the flux activity of three different pathways: vitamin A metabolism,

tetrahydrobiopterin metabolism and beta-alanine metabolism. The predictor split our population into high-risk and low-risk groups (p-value = 0.0032, HR=6.520, cut-off= 30%-70%) (Figure 2, Table 2).

Strikingly, the prognostic signature retained its prognostic value in the ER+ tumors, being the differences statistically significant in the ER-true group (p-value= 0.0179), but have worse performance in TNBC tumors (Figure 3).

Multivariate Cox regression model showed that this signature added information to clinical data (Sup Table 2).

Given that TNBC tumors have differences in metabolism when compared with ER+; a predictive signature only for TNBC was built. This signature was formed by the flux activities of glycolysis and glutamate metabolism and split TNBC population into a low- and a high-risk group (p-value= 0.1064, HR= 4.600, cut-off= 30-70%); however, it did not reach significance (Figure 4, Table 3).

In this case, multivariate analysis was not significant (Sup Table 3).

Discussion

Alterations in metabolism constitute a hallmark of cancer [3]. We previously described differences in metabolism between breast cancer subtypes [2,6]. TNBC and TN-like tumors had a higher growth rate than ER-true tumors in FBA predictions [2]. In the present work, our aim was to characterize additional differences in metabolic pathways between breast cancer subtypes.

The analysis of metabolism can use data coming both from genomics or proteomics. Proteomics provides more direct information about biological effectors than genomics, what is more useful in the GPR estimation. Proteomics experiments detected 3,239 proteins in the present study, although strict filtering criteria reduced the number to 1,095 proteins for analysis.

Comparing flux patterns in large cohorts is challenging. Flux activities are a method previously proposed to compare metabolic pathways between control cells and cells treated with drugs targeting metabolism [8]. Now, using flux activities we characterized differences in ROS detoxification and nucleotide interconversion pathways among breast tumor subtypes. Oxidative stress has been related with tumor aggressiveness in ER+ tumors [16]. TNBC tumors also have a high presence of ROS [17]. As expected, our model predicted lower ROS detoxification in ER-true tumors. Nucleotide interconversion category comprehends all the information about ATP and GTP metabolism. Differences in ATP or GTP production between breast cancer subtypes have not been previously described in the literature.

It was also possible to associate these flux activities with distant metastasis-free survival in this cohort. We found some flux activities related with distant relapse-free survival and built a predictor. The predictive signature included the flux activity of three metabolic pathways: vitamin A, tetrahydrobiopterin and beta-alanine metabolism. To our knowledge, this is the first time that data from FBA are associated with prognosis in cancer. Vitamin A or retinol has been related with risk of relapse in breast cancer [18], but there were no previous information about the prognostic value of tetrahydrobiopterin or beta-alanine metabolism. Additionally, the prognostic value of the predictor remained among ER+ tumors

It is well-known that TNBC tumors have differences in metabolism comparing with ER+ tumors [2,6]. For this reason, a predictor taking into account only TNBC tumors was done. This signature included flux activities of glycolysis and glutamate metabolism. However, survival

and multivariate analyses are not significant, probably due to the small number of samples. An increase in glucose uptake and lactate production associated to glycolysis was previously described in TNBC cells [6] It is also well-known that lactate dehydrogenase B and other glycolytic genes are upregulated in TNBC tumors as compared with the other subtypes and that LDHB overexpression is associated with poor clinical outcome [19]. On the other hand, TNBC tumors have a deregulation of glutaminolysis and exhibited the most frequent expression of proteins related with glutamine metabolism than other subtypes [20,21].

The study has some limitations. A validation of the predictors in an independent cohort is needed. Another limitation is the number of samples of each group in this cohort. It is also necessary to validate these findings into larger cohorts of patients. On the other hand, further improvements in mass-spectrometry techniques now allow the detection of more proteins, which would provide further information at GPR level.

Conclusions

In conclusion, flux activities, method proposed in previous works, now demonstrated its utility in summarizing FBA data and allows its association with prognosis. Differences in ROS detoxification and nucleotide interconversion pathways between breast cancer subtypes were characterized. Moreover, vitamin A, beta-alanine and tetrahydrobiopterin metabolism flux activities could be used to predict risk of distant metastasis-free survival in breast cancer patients. Finally, glycolysis and glutamate metabolism may be used to predict distant relapse-free survival in TNBC tumors.

Summary points

- Flux activities could be used to characterize differences in metabolic pathways between groups of tumors and to associate them with clinical outcomes.
- There are differences in flux activities of nucleotide interconversion and ROS detoxification between breast cancer subtypes.
- Flux activities of vitamin A, beta-alanine metabolism and tetrahydrobiopterin metabolism predict distant metastasis-free survival in breast cancer patients.
- Flux activities of glycolysis and glutamate metabolism may predict distant relapse-free survival in TNBC.

References

1. Siegel RL, Miller KD, Jemal A. Cancer statistics, 2018. *CA Cancer J Clin*, 68(1), 7-30 (2018).
2. Gámez-Pozo A, Trilla-Fuertes L, Berges-Soria J *et al.* Functional proteomics outlines the complexity of breast cancer molecular subtypes. *Scientific Reports*, 7(1), 10100 (2017).
3. Hanahan D, Weinberg RA. Hallmarks of cancer: the next generation. *Cell*, 144(5), 646-674 (2011).
4. Warburg O. The metabolism of carcinoma cells. *J Cancer Res*, 9, 148-163 (1925).
5. DeBerardinis RJ, Mancuso A, Daikhin E *et al.* Beyond aerobic glycolysis: transformed cells can engage in glutamine metabolism that exceeds the requirement for protein and nucleotide synthesis. *Proc Natl Acad Sci U S A*, 104(49), 19345-19350 (2007).
6. Gámez-Pozo A, Berges-Soria J, Arevalillo JM *et al.* Combined label-free quantitative proteomics and microRNA expression analysis of breast cancer unravel molecular differences with clinical implications. (Ed.^(Eds) (Cancer Res, 2015) 2243-2253.
7. Orth J, Thiele I, Palsson B. What is flux balance analysis? (Ed.^(Eds) (Nat Biotechnol, 2010) 245-248.
8. Trilla-Fuertes L, Gámez-Pozo A, Arevalillo JM *et al.* Molecular characterization of breast cancer cell response to metabolic drugs. *Oncotarget*, 9(11), 9645-9660 (2018).
9. Gámez-Pozo A, Ferrer NI, Ciruelos E *et al.* Shotgun proteomics of archival triple-negative breast cancer samples. *Proteomics Clin Appl*, 7(3-4), 283-291 (2013).
10. Gámez-Pozo A, Sánchez-Navarro I, Calvo E *et al.* Protein phosphorylation analysis in archival clinical cancer samples by shotgun and targeted proteomics approaches. *Mol Biosyst*, 7(8), 2368-2374 (2011).
11. Tyanova S, Temu T, Sinitcyn P *et al.* The Perseus computational platform for comprehensive analysis of (prote)omics data. *Nat Methods*, 13(9), 731-740 (2016).
12. Schellenberger J, Que R, Fleming R *et al.* Quantitative prediction of cellular metabolism with constraint-based models: the COBRA Toolbox v2.0. (Ed.^(Eds) (Nature Protocols, 2011) 1290-1307.
13. Thiele I, Swainston N, Fleming RM *et al.* A community-driven global reconstruction of human metabolism. *Nat Biotechnol*, 31(5), 419-425 (2013).
14. Barker BE, Sadagopan N, Wang Y *et al.* A robust and efficient method for estimating enzyme complex abundance and metabolic flux from expression data. *Comput Biol Chem*, 59 Pt B, 98-112 (2015).
15. Colijn C, Brandes A, Zucker J *et al.* Interpreting expression data with metabolic flux models: Predicting Mycobacterium tuberculosis mycolic acid production. (Ed.^(Eds) (PLOS Comput Bio, 2009)
16. Mahalingaiah PK, Ponnusamy L, Singh KP. Chronic oxidative stress causes estrogen-independent aggressive phenotype, and epigenetic inactivation of estrogen receptor alpha in MCF-7 breast cancer cells. *Breast Cancer Res Treat*, 153(1), 41-56 (2015).
17. Pelicano H, Zhang W, Liu J *et al.* Mitochondrial dysfunction in some triple-negative breast cancer cell lines: role of mTOR pathway and therapeutic potential. *Breast Cancer Res*, 16(5), 434 (2014).
18. Formelli F, Meneghini E, Cavadini E *et al.* Plasma retinol and prognosis of postmenopausal breast cancer patients. *Cancer Epidemiol Biomarkers Prev*, 18(1), 42-48 (2009).
19. McCleland ML, Adler AS, Shang Y *et al.* An integrated genomic screen identifies LDHB as an essential gene for triple-negative breast cancer. *Cancer Res*, 72(22), 5812-5823 (2012).

20. Kim S, Kim DH, Jung WH, Koo JS. Expression of glutamine metabolism-related proteins according to molecular subtype of breast cancer. *Endocr Relat Cancer*, 20(3), 339-348 (2013).
21. Lampa M, Arlt H, He T *et al.* Glutaminase is essential for the growth of triple-negative breast cancer cells with a deregulated glutamine metabolism pathway and its suppression synergizes with mTOR inhibition. *PLoS One*, 12(9), e0185092 (2017).

Financial disclosure: This work was supported by Instituto de Salud Carlos III, Spanish Economy and Competitiveness Ministry, Spain and co-funded by the FEDER program, "Una forma de hacer Europa" [PI15/01310]; LT-F is supported by the Spanish Economy and Competitiveness Ministry [DI-15-07614]; GP-V is supported by the Consejería de Educación, Juventud y Deporte of Comunidad de Madrid [IND2017/BMD7783]; AZ-M is supported by Jesús Antolín Garciarena fellowship from IdiPAZ. The funders had no role in the study design, data collection and analysis, decision to publish or preparation of the manuscript.

Conflicts of interest: JAFV, EE and AG-P are shareholders in Biomedica Molecular Medicine SL. LT-F and GP-V are employees of Biomedica Molecular Medicine SL. The other authors declare that they have no competing interests.

Ethical conduct: The study was approved by the Ethical Committee of Hospital 12 de Octubre and Hospital La Paz.

Figure and table legends

Table 1: Patient characteristics.

Table 2: Weights assigned to each metabolic pathway contained in the predictor signature for all tumors.

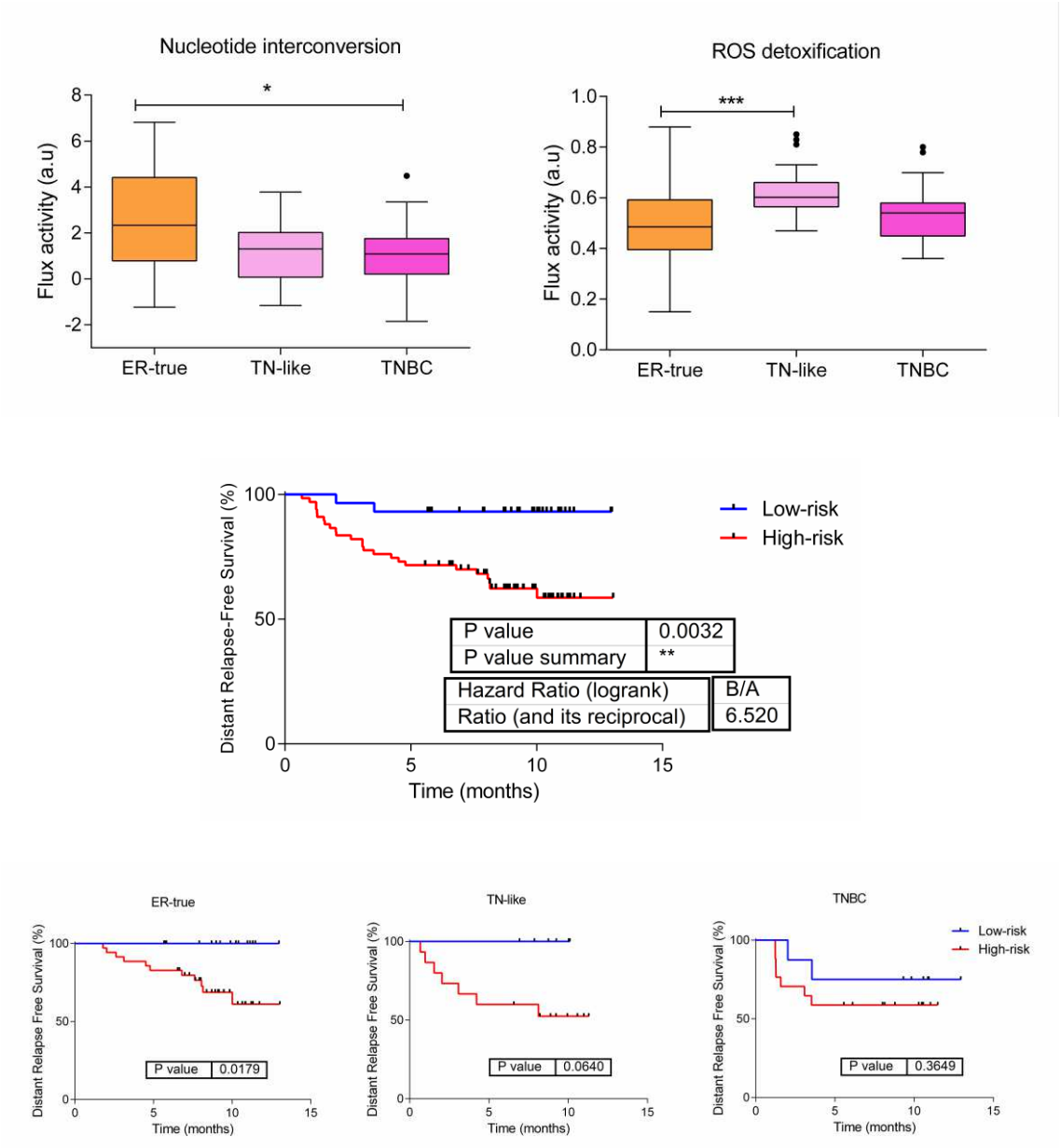
Table 3: Weights assigned to each metabolic pathway contained in the predictor signature for TNBC tumors.

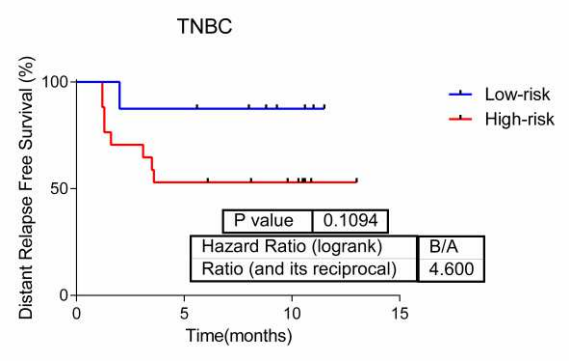
Figure 1: Flux activities with significant differences between breast cancer subtypes (* = 0.01 to 0.05, significant; *** = 0.0001 to 0.001, extremely significant).

Figure 2: Predictive signature based on the flux activities of vitamin A metabolism, tetrahydrobiopterin metabolism and beta-alanine metabolism.

Figure 3: Predictor based on beta-alanine, tetrahydrobiopterin and vitamin A flux activities in the different breast cancer subtypes.

Figure 4: Predictor of distant relapse-free survival in TNBC tumors based on the flux activities of glycolysis and glutamate metabolism.





Probabilistic graphical models and computational metabolic models applied to the analysis of metabolomics data in breast cancer

Lucía Trilla-Fuertes[1¶]; Angelo Gámez-Pozo[1,2¶]; Jorge M Arevalillo[3]; Guillermo Prado-Vázquez[1]; Andrea Zapater-Moros[1,2]; Mariana Díaz-Almirón[4]; Hilario Navarro[3]; Paloma Maín[5]; Enrique Espinosa[6,7]; Pilar Zamora[6,7]; and Juan Ángel Fresno Vara[2,7,*]

[1]Biomedica Molecular Medicine SL, Madrid, Spain

[2]Molecular Oncology & Pathology Lab, Institute of Medical and Molecular Genetics-INGEMM, La Paz University Hospital-IdiPAZ, Madrid, Spain

[3]Department of Statistics, Operational Research and Numerical Analysis, National University of Distance Education (UNED).

[4]Biostatistics Unit, La Paz University Hospital-IdiPAZ, Madrid, Spain

[5]Department of Statistics and Operations Research, Faculty of Mathematics, Complutense University of Madrid, Madrid, Spain.

[6]Medical Oncology Service, La Paz University Hospital-IdiPAZ, Madrid, Spain

[7]Biomedical Research Networking Center on Oncology-CIBERONC, ISCIII, Madrid, Spain

[¶] These authors contributed equally to this work.

[*] Corresponding Author: juanangel.fresno@salud.madrid.org (JAFV)

Short title: Computational models applied to metabolomics data in breast cancer

Abstract

Metabolomics has great potential in the development of new biomarkers in cancer. In this study, metabolomics and gene expression data from breast cancer tumor samples were analyzed, using (1) probabilistic graphical models to define associations using quantitative data without other *a priori* information; and (2) Flux Balance Analysis and flux activities to characterize differences in metabolic pathways. A metabolite network was built through the use of probabilistic graphical models. Interestingly, the metabolites were organized into metabolic pathways in this network, thus it was possible to establish differences between breast cancer subtypes at the metabolic pathway level. Additionally, the lipid metabolism node had prognostic value. A second network associating gene expression with metabolites was built. Associations were established between the biological functions of genes and the metabolites included in each node. A third network combined flux activities from Flux Balance Analysis and metabolomics data, showing coherence between the metabolic pathways of the flux activities and the metabolites in each branch. In this study, probabilistic graphical models were valuable for the functional analysis of metabolomics data from a functional point of view, allowing new hypotheses in metabolomics and associating metabolomics data with the patient's clinical outcome.

Author summary

Metabolomics is a promising technique to describe new biomarkers in cancer. In this study we proposed computational methods to manage this type of data and associate it with gene expression data. We also employed a metabolic computational model to compare predictions from this model with metabolomics measurements. Finally, we built predictors of relapse based on the integration of those high-dimensional data in breast cancer patients.

Introduction

Breast cancer is one of the most common malignancies, with 266,120 estimated new cases and 40,920 estimated deaths in the United States in 2018 [1]. In clinical practice, the expression of hormonal receptors and HER2 allows the classification of this disease into three groups: hormonal receptor-positive (ER+), HER2+ and triple negative (TNBC).

Metabolomics is the most recent -omics. It consists of measuring the entire set of metabolites present in a biological sample [2]. The most common techniques in metabolomics experiments are mass spectrometry-related methods, which are based on the mass/charge relationships of each metabolite or its fragments [3]. Metabolomics is a promising tool for the development of new biomarkers [4].

We used two different methods to merge metabolomics and gene expression data in breast cancer. In previous studies, we used probabilistic graphical models (PGMs) to study differences between breast tumor subtypes and to characterize muscle-invasive bladder cancer at a functional level using proteomics data [5-7]. Flux Balance Analysis (FBA), however, is a method that has been widely used to study biochemical networks [8]. FBA predicts the growth rate or the rate of production of a given metabolite [9], and it has previously been used to characterize breast cancer cell responses against drugs targeting metabolism [10]. In this study, flux activities were proposed as a feasible method to compare flux patterns in metabolic pathways.

In the present study, metabolomics and gene expression data from 67 fresh tissue samples [11] were analyzed through PGMs and FBA. Our aim was to find associations between metabolomics and gene expression data.

Results

Patient characteristics

The data used in this study are from the previous work of Terunuma et al. [11]. A total of 67 paired normal and tumor fresh tissue samples from patients with breast cancer were studied. We only selected samples from tumor tissues for the present analyses.

This cohort included 67 patients, 33 ER+ and 34 ER- (of which 14 were TNBC). The median follow-up was 50 months, and 31 deaths had occurred during this time. No significant differences regarding overall survival were observed between patients with ER+ or ER- tumors. Patient characteristics are shown in Table 1.

	n (%)	ER+	ER-
Number of patients	67	33	34
Age (years)			
Median	51	57	48
Range	30–93	34–93	30–75
TNM stage			
I	6 (9%)	4 (12%)	2 (6%)
II	2 (3%)	1 (3%)	1 (3%)
IIA	23 (35%)	12 (37%)	11 (32%)
IIB	21 (31%)	7 (21%)	14 (41%)
IIIA	9 (13%)	5 (15%)	4 (12%)
IIIB	6 (9%)	4 (12%)	2 (6%)
N category			
pN0	37 (55%)	17 (52%)	20 (59%)

pN1	24 (35%)	13 (39%)	11 (32%)
pN2	5 (8%)	3 (9%)	2 (6%)
Missing	1 (2%)	0 (0%)	1 (3%)
Grade			
G1	8 (12%)	8 (24%)	0 (0%)
G2	20 (30%)	14 (43%)	6 (18%)
G3	29 (43%)	7 (21%)	22 (64%)
Missing	10 (15%)	4 (12%)	6 (18%)
Neoadjuvant therapy			
Yes	6 (9%)	2	4 (12%)
No	50 (75%)	26	24 (70%)
Missing	11 (16%)	5	6 (18%)

Table 1: Patient characteristics.

Analysis of metabolomics data

An overall survival predictor using metabolomics data was built. This signature included five metabolites: glutamine, 2-hydroxypalmitate, deoxycarnitine, butyrylcarnitine and glycerophosphorylcholine (p-value =0.003, hazard ratio [HR] = 0.342, cut-off = 50:50) (Fig 1). A multivariate analysis showed that the predictor provided additional prognostic information to that of the clinical data (S1 Table).

Fig 1: Predictive signature built using metabolomics data.

Metabolomics data, including 237 metabolites, were analyzed through PGM. The resulting network was built assigning a main metabolic pathway to each node using IMPaLA. IMPaLA is a tool that allows ontology analyses based on metabolic pathways instead of genes. Strikingly,

this network had a functional structure, grouping the metabolites into metabolic pathways. The network had five nodes, each with a different overrepresented metabolic pathway (Fig 2).

Fig 2: Probabilistic graphical model from metabolomics data.

The activity of each node was calculated as previously described [6, 7, 10, 12]. Significant differences were found between ER+ and ER- tumors in lipid metabolism and purine metabolism ($p<0.05$) (S1 Fig).

The lipid metabolism node had prognostic value in this cohort ($p = 0.045$, HR = 0.479, cut-off = 50:50) (Fig 3). Differences remained when stratified by the expression of hormonal receptors. However, a multivariate analysis did not show that the predictor supplied additional prognostic information to that of the clinical data (S2 Table).

Fig 3: Predictor based on lipid metabolism node activity.

Analyses combining gene expression with metabolomics data

A network combining metabolomics and gene expression data was built. Although most metabolites were grouped in the same node, some were integrated into gene nodes (Fig 4).

Fig 4: A. Network associating genes (red) and metabolites (blue). B. Metabolite and gene network functionally characterized.

This combined network was then functionally characterized. The resulting network had eleven functional nodes and a twelfth node that grouped the metabolites (Fig 4).

Once the main functions were assigned, a literature review was performed to study the relationship between metabolites included in the gene nodes and the main function of each node. A relationship with functional nodes had been previously described for 4 of 20

metabolites: succinate, cytidine, histamine and 1,2-propanediol. The relationships between metabolites and their node function are shown in Table 2.

Metabolite	Node	Described relationship	Reference
Succinate	Immune response	Increases immune response, induces IL-1b production, promotes adaptive immune response.	PMID: 28109906
Cytidine	Immune response	5-aza-2'-deoxycytidine potentiates antitumor immune response, role in innate immune response	PMID: 23865062, PMID: 24559534
Histamine	Angiogenesis	Histamine promotes angiogenesis by enhancing VEGF production	PMID: 23225320
1,2-propanediol (prev.X-4796)	Angiogenesis	Modulates the immune system through S1P, which promotes angiogenesis and proliferation. 14C-sulfoquinovosyl acylpropanediol is an antiangiogenic drug	PMID: 21632869, PMID: 29543539

Table 2: Previously described relationships between metabolites included in gene nodes and the function of these nodes.

Flux Balance Analysis and flux activities

FBA and flux activities were calculated as previously described [10]. No significant differences were found in the tumor growth rate between ER+ and ER- tumors (S2 Fig).

Flux activities showed significant differences between ER+ and ER- in glycerophospholipid metabolism, phosphatidyl inositol metabolism, urea cycle, propanoate metabolism, pyrimidine catabolism and reactive oxygen species (ROS) detoxification (S3 Fig).

A predictor for overall survival was built with flux activities of glutamate metabolism and alanine and aspartate metabolism (p-value = 0.024, HR = 0.411, cut-off = 50:50) (Fig 5). A multivariate analysis showed that the predictor provided prognostic information independent from clinical data (S3 Table).

Fig 5: OS predictor based on glutamate metabolism and alanine and aspartate metabolism flux activities.

PGM analysis combining flux activities with metabolomics data

Using flux activities and metabolomics data, a new network was built. Interestingly, this network combined both types of data; however, flux activities appeared at the periphery of the network (Fig 6).

Fig 6: A. Network combining flux activities (purple) and metabolite (pink) expression data. B. Division in branches of the network formed by flux activities and metabolomics data.

The resulting network was split into several branches to study the relationship of the metabolites to the flux activities included in each branch (Fig 6). Coherence between both types of data was shown, associating flux activities and metabolites related to these flux activities in the same branch. For instance, branch 1 includes glycolysis flux activity and three metabolites previously related to glycolysis (S4 Table). Regarding vitamins and cofactors, it was not possible make comparisons because the IMPaLA label for this category is “Vitamin and co-factor metabolism” and Recon2 labels differentiate between the various vitamins, labeling them as “Vitamin B6 metabolism”, “Vitamin A metabolism”, etc.

Discussion

Metabolomics is attracting considerable interest as a technique for finding new biomarkers in cancer. In this study, a new analytical workflow for the management and study of metabolomics data was proposed. This workflow allowed global metabolic characterization, beyond analyses based on unique metabolites.

Genomics and metabolomics data from Terunuma et al. have previously been used by The Cancer Genome Atlas Consortium to correlate gene expression data with metabolomics data [11, 13]. Based on this dataset, we applied PGMs for the first time in metabolomics data from tumor samples and also in metabolomics data combined with gene expression data and flux activities, with the aim of confirming known associations and finding new ones.

First, we evaluated whether metabolomics data were related to overall survival in patients with breast cancer. An overall survival predictive signature was built that included the expression values of glutamine, deoxycarnitine, butyrylcarnitine, glycerophosphorylcholine and 2-hydroxypalmitate [14]. The first three of these metabolites has previously been related to survival in breast cancer [15, 16]. However, to our knowledge, this is the first report associating 2-hydroxypalmitate with cancer survival. Additionally, in the previous study by Terunuma et al., 2-hydroxyglutarate was associated with a poor prognosis in patients with breast cancer [11]. 2-hydroxyglutarate is a glutamine intermediate in the tricarboxylic acid cycle, involved in the conversion of glutamine into lactate, a process known as glutaminolysis [14]. These results highlight the relevance of glutamine metabolism in breast cancer prognoses.

A metabolite network using metabolomics data was built using PGM. IMPaLA assigned a dominant metabolic function to each resulting node. In previous studies, we demonstrated that PGMs are useful for functionally characterizing gene or protein networks [6, 7, 12].

However, to our knowledge, this is the first time a PGM has been applied to metabolomics data from tumor samples. Just as observed in genes or proteins, metabolites are grouped into metabolic pathways, allowing the characterization of differences in metabolic pathways between ER+ and ER- tumors. For example, both lipid metabolism and purine metabolism node activities were higher in ER- tumors. ER- tumors usually overexpress genes related to lipid metabolism. [17]. Moreover, the activity of the lipid metabolism node had prognostic value. No relationship between purine metabolism and breast cancer has previously been defined.

On the other hand, the network combining gene expression data and metabolomics data grouped most of the metabolites into an isolated node. Yet, some metabolites were included in gene nodes. We found that four of the twenty metabolites showed a previously reported relationship with the main function of the gene node in which they were included. Succinate and cytidine were located in the immune response node. Succinate acts as an inflammation activation signal, inducing IL-1β cytokine production through hypoxia-inducible factor 1 [18]. In addition, succinate increases dendritic cell capability to act as antigen-presenting cells, prompting an adaptive immune response [19]. Regarding cytidine, Wachowska et al. described that 5-aza-2'-deoxycytidine modulates the levels of major histocompatibility complex class I molecules in tumor cells, induces P1A antigen and has immunomodulatory activity when combined with photodynamic therapy [20].

Both histamine and 1,2-propanediol appeared to be related to the angiogenesis node. Histamine is known to promote angiogenesis through vascular epithelial growth factor [21]. On the other hand, sulfoquinovosyl acylpropanediol, an 1,2-propanediol derivate, inhibits angiogenesis in murine models with pulmonary carcinoma [22].

FBA was used to model metabolism using gene expression data. Although FBA-predicted biomass did not show significant differences between ER+ and ER- tumors, differences in flux activities were shown between both subtypes. Some of these activities were also related to prognosis. One of these flux activities is “Glutamine metabolism”, which agrees with the results obtained from the metabolomics data, including glutamine in the metabolite, a signature capable of predicting overall survival. With the aim of associating metabolomics and FBA results, flux activities and metabolomics data were combined to form a new network. As opposed to gene and metabolite data, metabolomics data and flux activities combined well in the network. Interestingly, flux activities are dead-end nodes, perhaps due to the fact that they are by definition a final summary of each pathway. IMPaLA assigned a main metabolic pathway to resulting branches; thus, it was possible to know how many metabolites were related to flux activity in each branch. In most cases with available information, there was coherence between metabolites included in the branch and its flux activity. This validates FBA and flux activities, both based on gene expression, as a method of simulating metabolism.

Our study has some limitations. The limited number of samples leads us to consider the results as preliminary, and validation in an independent cohort is needed. Additionally, our results are difficult to place in the current clinical landscape, given that tumors in the original series had not been assessed for HER2 expression. On the other hand, evolving techniques currently allow the detection of more metabolites, which would permit a more thorough analysis.

In conclusion, PGMs reveal their utility in the analysis of metabolomics data from a functional point of view, not only metabolomics data alone, but also in combination with flux or gene expression data. Therefore, PGM is postulated as a method to propose new hypotheses in the metabolomics field. We also found that it is possible to associate metabolomics data with clinical outcomes and to build prognostic signatures based on metabolomics data.

Materials and methods

Patients included in the study

Metabolomics and gene expression data from 67 fresh tumor tissue samples originally analyzed by Terunuma et al. [11] were included in this study.

Preprocessing of gene expression and metabolomics data

For the metabolomics data, log2 was calculated. As quality criteria, data were filtered to include detectable measurements in at least 75% of the samples. Missing values were imputed to a normal distribution using Perseus software [23]. After quality control, 237 metabolites were considered for subsequent analyses.

In terms of gene expression data, the 2000 most variable genes, i.e., those genes with the highest standard deviation, were chosen to build the PGM.

Probabilistic graphical models and gene ontology analyses

As previously described [6, 7, 10, 12], PGMs compatible with high dimensional data were used, using correlations as associative criteria. The *grapHD* package [24] and R v3.2.5 [25] were employed to build the network. A majority function was assigned to each node using gene ontology analyses. In the case of genes, gene ontology analyses were performed using the DAVID web tool with “homo sapiens” as background and GOTERM, KEGG and Biocarta selected as categories [26]. In the case of metabolites, the Integrated Molecular Pathway Level Analysis (IMPaLA) web tool was used [27].

Node activities were calculated, as previously described [6, 7, 10, 12], as the mean of the expression/quantity of genes/metabolites of each node that are related to the main node function/metabolic pathway.

Flux Balance Analysis and flux activities

FBA was calculated using the human metabolic reconstruction Recon2 [28]. As the objective function, the biomass reaction proposed in the Recon2 was used. FBA was performed using the COBRA Toolbox [29] available for MATLAB. Gene-Protein-Reaction rules were solved as described in previous studies [7, 10], using a modification of the Barker *et al.* algorithm [30], which were incorporated into the model by the E-flux method [31].

Flux activities were previously proposed as a measurement to compare differences at the flux pathway level [10]. Briefly, they were calculated as the sum of the fluxes of the reactions included in each pathway defined in Recon2.

Statistical analyses

The statistical analyses were performed with GraphPad Prism v6, and the network analyses were performed using Cytoscape software [32]. Predictor signatures were built with the BRB Array Tool from Dr. Richard Simon's team. All p-values are two-sided and are considered statistically significant under 0.05.

Funding statement

This study was supported by Instituto de Salud Carlos III, Spanish Economy and Competitiveness Ministry, Spain and co-funded by the FEDER program, "Una forma de hacer Europa" (PI15/01310). LT-F is supported by the Spanish Economy and Competitiveness Ministry (DI-15-07614). GP-V is supported by Conserjería de Educación, Juventud y Deporte of Comunidad de Madrid (IND2017/BMD7783). The funders had no role in the study design, data collection and analysis, decision to publish or preparation of the manuscript.

Author contributions

All the authors have directly participated in the preparation of this manuscript and have approved the final version submitted. JMA, MD-A, HN and PM contributed the directed graphical models. LT-F, AG-P, G-PV, and AZ-M performed the statistical analyses, the graphical model interpretation and the ontology analyses. LT-F, AG-P, JAFV, PZ and EE conceived of the study and participated in its design and interpretation. MD-A and LT-F performed the Flux Balance Analysis. LT-F drafted the manuscript. AG-P, JAFV, and EE supported the manuscript drafting. AG-P and JAFV coordinated the study.

Competing interests

JAFV, EE and AG-P are shareholders in Biomedica Molecular Medicine SL. LT-F and GP-V are employees of Biomedica Molecular Medicine SL. The other authors declare no competing interests.

References

1. Siegel RL, Miller KD, Jemal A. Cancer statistics, 2018. CA Cancer J Clin. 2018;68(1):7-30. Epub 2018/01/04. doi: 10.3322/caac.21442. PubMed PMID: 29313949.
2. Fiehn O. Metabolomics--the link between genotypes and phenotypes. Plant Mol Biol. 2002;48(1-2):155-71. PubMed PMID: 11860207.
3. Emwas AH. The strengths and weaknesses of NMR spectroscopy and mass spectrometry with particular focus on metabolomics research. Methods Mol Biol. 2015;1277:161-93. doi: 10.1007/978-1-4939-2377-9_13. PubMed PMID: 25677154.
4. Gowda GA, Zhang S, Gu H, Asiago V, Shanaiah N, Raftery D. Metabolomics-based methods for early disease diagnostics. Expert Rev Mol Diagn. 2008;8(5):617-33. doi: 10.1586/14737159.8.5.617. PubMed PMID: 18785810; PubMed Central PMCID: PMCPMC3890417.
5. Sánchez-Navarro I, Gámez-Pozo A, Pinto A, Hardisson D, Madero R, López R, et al. An 8-gene qRT-PCR-based gene expression score that has prognostic value in early breast cancer. BMC Cancer. 2010;10:336. Epub 2010/06/28. doi: 10.1186/1471-2407-10-336. PubMed PMID: 20584321; PubMed Central PMCID: PMCPMC2906483.
6. Gámez-Pozo A, Berges-Soria J, Arevalillo JM, Nanni P, López-Vacas R, Navarro H, et al. Combined label-free quantitative proteomics and microRNA expression analysis of breast cancer unravel molecular differences with clinical implications. Cancer Res; 2015. p. 2243-53.
7. Gámez-Pozo A, Trilla-Fuertes L, Berges-Soria J, Selevsek N, López-Vacas R, Díaz-Almirón M, et al. Functional proteomics outlines the complexity of breast cancer molecular subtypes. Scientific Reports. 2017;7(1):10100. doi: 10.1038/s41598-017-10493-w.
8. Varma A, Palsson BO. Parametric sensitivity of stoichiometric flux balance models applied to wild-type Escherichia coli metabolism. Biotechnol Bioeng. 1995;45(1):69-79. doi: 10.1002/bit.260450110. PubMed PMID: 18623053.
9. Orth J, Thiele I, Palsson B. What is flux balance analysis? : Nat Biotechnol; 2010. p. 245-8.
10. Trilla-Fuertes L, Gámez-Pozo A, Arevalillo JM, Díaz-Almirón M, Prado-Vázquez G, Zapater-Moros A, et al. Molecular characterization of breast cancer cell response to metabolic drugs. Oncotarget. 2018;9(11):9645-60. Epub 2018/01/08. doi: 10.18632/oncotarget.24047. PubMed PMID: 29515760; PubMed Central PMCID: PMCPMC5839391.
11. Terunuma A, Putluri N, Mishra P, Mathé EA, Dorsey TH, Yi M, et al. MYC-driven accumulation of 2-hydroxyglutarate is associated with breast cancer prognosis. J Clin Invest. 2014;124(1):398-412. Epub 2013/12/09. doi: 10.1172/JCI71180. PubMed PMID: 24316975; PubMed Central PMCID: PMCPMC3871244.
12. de Velasco G, Trilla-Fuertes L, Gamez-Pozo A, Urbanowicz M, Ruiz-Ares G, Sepúlveda JM, et al. Urothelial cancer proteomics provides both prognostic and functional information. Sci Rep. 2017;7(1):15819. Epub 2017/11/17. doi: 10.1038/s41598-017-15920-6. PubMed PMID: 29150671; PubMed Central PMCID: PMCPMC5694001.
13. Peng X, Chen Z, Farshidfar F, Xu X, Lorenzi PL, Wang Y, et al. Molecular Characterization and Clinical Relevance of Metabolic Expression Subtypes in Human Cancers. Cell Rep. 2018;23(1):255-69.e4. doi: 10.1016/j.celrep.2018.03.077. PubMed PMID: 29617665; PubMed Central PMCID: PMCPMC5916795.
14. DeBerardinis RJ, Mancuso A, Daikhin E, Nissim I, Yudkoff M, Wehrli S, et al. Beyond aerobic glycolysis: transformed cells can engage in glutamine metabolism that exceeds the requirement for protein and nucleotide synthesis. Proc Natl Acad Sci U S A. 2007;104(49):19345-50. doi: 10.1073/pnas.0709747104. PubMed PMID: 18032601; PubMed Central PMCID: PMCPMC2148292.

15. Bhowmik SK, Ramirez-Peña E, Arnold JM, Putluri V, Sphyris N, Michailidis G, et al. EMT-induced metabolite signature identifies poor clinical outcome. Oncotarget. 2015;6(40):42651-60. doi: 10.18632/oncotarget.4765. PubMed PMID: 26315396; PubMed Central PMCID: PMCPMC4767460.
16. Cao MD, Sitter B, Bathen TF, Bofin A, Lønning PE, Lundgren S, et al. Predicting long-term survival and treatment response in breast cancer patients receiving neoadjuvant chemotherapy by MR metabolic profiling. NMR Biomed. 2012;25(2):369-78. Epub 2011/08/08. doi: 10.1002/nbm.1762. PubMed PMID: 21823183.
17. Wang J, Shidfar A, Ivancic D, Ranjan M, Liu L, Choi MR, et al. Overexpression of lipid metabolism genes and PBX1 in the contralateral breasts of women with estrogen receptor-negative breast cancer. Int J Cancer. 2017;140(11):2484-97. Epub 2017/03/21. doi: 10.1002/ijc.30680. PubMed PMID: 28263391.
18. Tannahill GM, Curtis AM, Adamik J, Palsson-McDermott EM, McGettrick AF, Goel G, et al. Succinate is an inflammatory signal that induces IL-1β through HIF-1α. Nature. 2013;496(7444):238-42. Epub 2013/03/24. doi: 10.1038/nature11986. PubMed PMID: 23535595; PubMed Central PMCID: PMCPMC4031686.
19. Jiang S, Yan W. Succinate in the cancer-immune cycle. Cancer Lett. 2017;390:45-7. Epub 2017/01/18. doi: 10.1016/j.canlet.2017.01.019. PubMed PMID: 28109906.
20. Wachowska M, Gabrysiak M, Muchowicz A, Bednarek W, Barankiewicz J, Rygiel T, et al. 5-Aza-2'-deoxycytidine potentiates antitumour immune response induced by photodynamic therapy. Eur J Cancer. 2014;50(7):1370-81. Epub 2014/02/18. doi: 10.1016/j.ejca.2014.01.017. PubMed PMID: 24559534; PubMed Central PMCID: PMCPMC4136636.
21. Lu Q, Wang C, Pan R, Gao X, Wei Z, Xia Y, et al. Histamine synergistically promotes bFGF-induced angiogenesis by enhancing VEGF production via H1 receptor. J Cell Biochem. 2013;114(5):1009-19. doi: 10.1002/jcb.24440. PubMed PMID: 23225320.
22. Ruike T, Kanai Y, Iwabata K, Matsumoto Y, Murata H, Ishima M, et al. Distribution and metabolism of. Xenobiotica. 2018:1-45. Epub 2018/03/15. doi: 10.1080/00498254.2018.1448949. PubMed PMID: 29543539.
23. Tyanova S, Temu T, Sinitcyn P, Carlson A, Hein MY, Geiger T, et al. The Perseus computational platform for comprehensive analysis of (prote)omics data. Nat Methods. 2016;13(9):731-40. Epub 2016/06/27. doi: 10.1038/nmeth.3901. PubMed PMID: 27348712.
24. Abreu G, Edwards D, Labouriau R. High-Dimensional Graphical Model Search with the gRapHD R Package Journal of Statistical Software2010. p. 1-18.
25. R Core Team. R: A language and environment for statistical computing. Vienna,Austria. R Foundation for Stattistical Computing, 2013.
26. Huang dW, Sherman BT, Lempicki RA. Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. Nat Protoc. 2009;4(1):44-57. doi: 10.1038/nprot.2008.211. PubMed PMID: 19131956.
27. Cavill R, Kamburov A, Ellis JK, Athersuch TJ, Blagrove MS, Herwig R, et al. Consensus-phenotype integration of transcriptomic and metabolomic data implies a role for metabolism in the chemosensitivity of tumour cells. PLoS Comput Biol. 2011;7(3):e1001113. Epub 2011/03/31. doi: 10.1371/journal.pcbi.1001113. PubMed PMID: 21483477; PubMed Central PMCID: PMCPMC3068923.
28. Thiele I, Swainston N, Fleming RM, Hoppe A, Sahoo S, Aurich MK, et al. A community-driven global reconstruction of human metabolism. Nat Biotechnol. 2013;31(5):419-25. Epub 2013/03/03. doi: 10.1038/nbt.2488. PubMed PMID: 23455439; PubMed Central PMCID: PMCPMC3856361.
29. Schellenberger J, Que R, Fleming R, Thiele I, Orth J, Feist A, et al. Quantitative prediction of cellular metabolism with constraint-based models: the COBRA Toolbox v2.0. Nature Protocols; 2011. p. 1290-307.

30. Barker BE, Sadagopan N, Wang Y, Smallbone K, Myers CR, Xi H, et al. A robust and efficient method for estimating enzyme complex abundance and metabolic flux from expression data. Comput Biol Chem. 2015;59 Pt B:98-112. Epub 2015/09/01. doi: 10.1016/j.compbiolchem.2015.08.002. PubMed PMID: 26381164; PubMed Central PMCID: PMCPMC4684447.

31. Colijn C, Brandes A, Zucker J, Lun D, Weiner B, Farhat M, et al. Interpreting expression data with metabolic flux models: Predicting Mycobacterium tuberculosis mycolic acid production. PLOS Comput Bio; 2009.

32. Shannon P, Markiel A, Ozier O, Baliga NS, Wang JT, Ramage D, et al. Cytoscape: a software environment for integrated models of biomolecular interaction networks. Genome Res. 2003;13(11):2498-504. doi: 10.1101/gr.1239303. PubMed PMID: 14597658; PubMed Central PMCID: PMCPMC403769.

Supporting information

S1 Table: Multivariate Cox regression model comparing OS predictor based on metabolomics data. T = tumor stage, N = lymph node status, G = tumor grade.

S2 Table: Multivariate Cox analysis comparing predictor based on node activity of lipid metabolism. T = tumor stage, N = lymph node status, G = tumor grade.

S3 Table: Multivariate Cox regression comparing predictor based on flux activities. T = tumor stage, N = lymph node status, G = tumor grade.

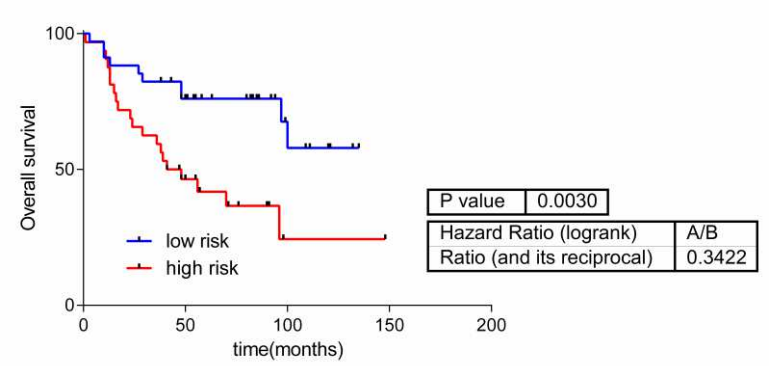
S4 Table: Metabolites associated with flux activity of each network branch.

S1 Fig: Node activities from the metabolic network.

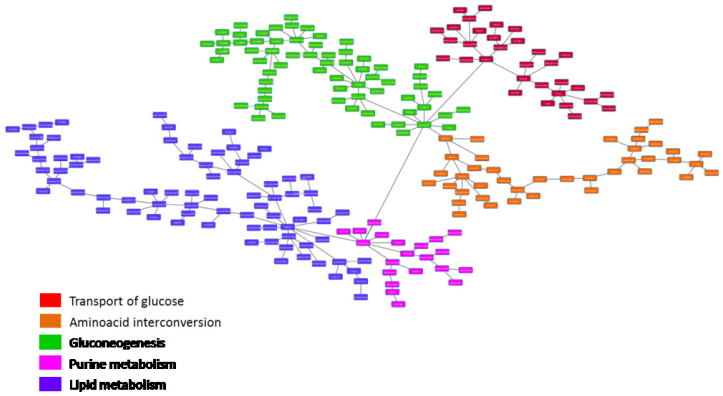
S2 Fig: Tumor growth rate predicted using FBA for ER+ and ER- tumors.

S3 Fig: Flux activities were significantly different between ER+ and ER-.

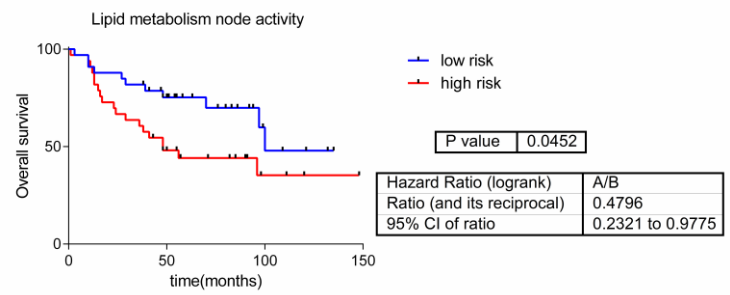
388



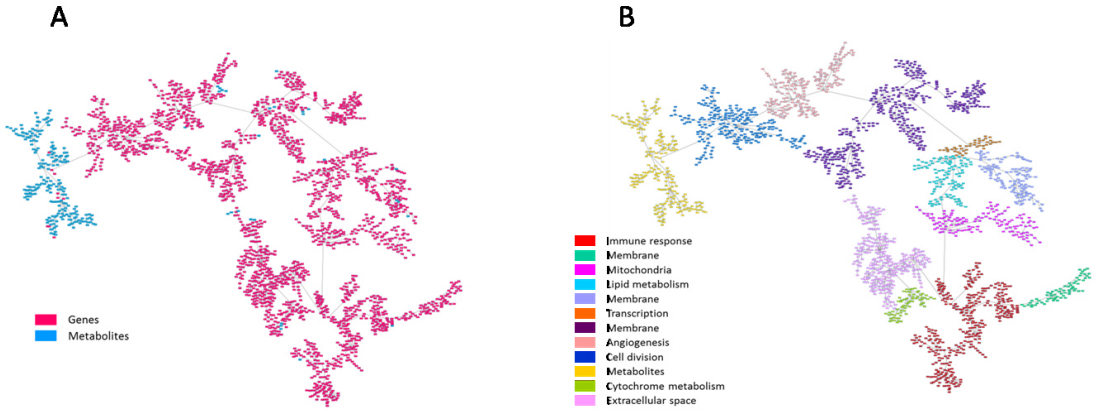
389

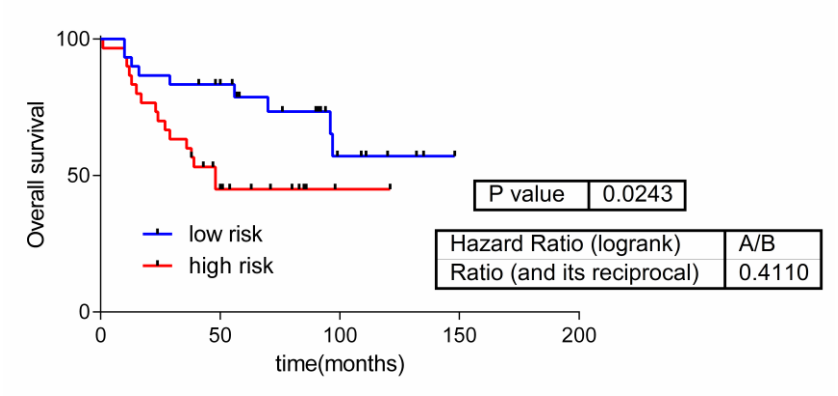


390

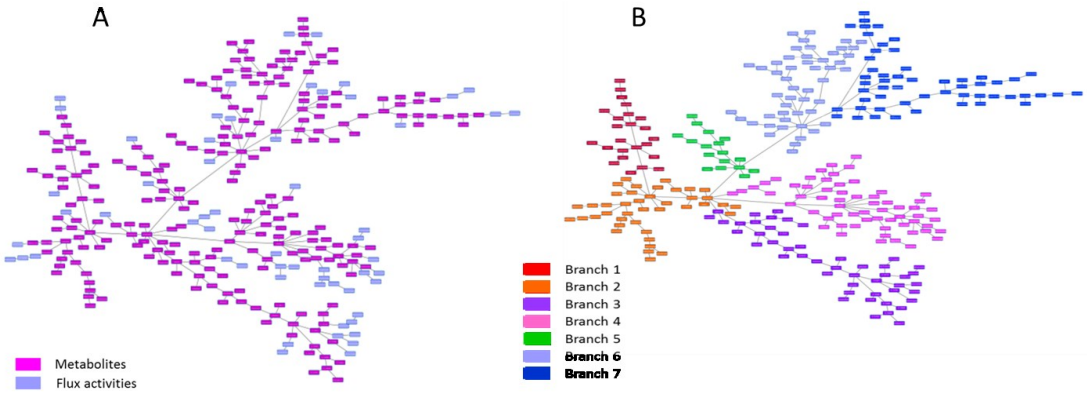


391





392



393

PLOS ONE

RESEARCH ARTICLE

Prediction of adjuvant chemotherapy response in triple negative breast cancer with discovery and targeted proteomics

Angelo Gámez-Pozo[1,2☯], Lucía Trilla-Fuertes[2☯], Guillermo Prado-Vázquez[1], Cristina Chiva[3,4], Rocío López-Vacas[1], Paolo Nanni[5], Julia Berges-Soria[1], Jonas Grossmann[5], Mariana Díaz-Almirón[6], Eva Ciruelos[7], Eduard Sabidó[3,4], Enrique Espinosa[8,9], Juan Ángel Fresno Vara[1,2,9]*

1 Molecular Oncology & Pathology Lab, Instituto de Genética Médica y Molecular-INGEMM, Hospital Universitario La Paz-IdiPAZ, Madrid, Spain, **2** Biomedica Molecular Medicine SL, Madrid, Spain, **3** Proteomics Unit, Center of Genomics Regulation (CRG), Barcelona Institute of Science and Technology (BIST), Barcelona, Spain, **4** Proteomics Unit, Universitat Pompeu Fabra (UPF), Barcelona, Spain, **5** Functional Genomics Centre Zurich, University of Zurich/ETH Zurich, Zurich, Switzerland, **6** Biostatistics Unit, Hospital Universitario La Paz-IdiPAZ, Madrid, Spain, **7** Medical Oncology Service, Instituto de Investigación Hospital Universitario Doce de Octubre-i+12, Madrid, Spain, **8** Medical Oncology Service, Hospital Universitario La Paz-IdiPAZ, Madrid, Spain, **9** CIBERONC. Instituto de Salud Carlos III, Madrid, Spain

☯ These authors contributed equally to this work.
* juanangel.fresno@salud.madrid.org

Citation: Gámez-Pozo A, Trilla-Fuertes L, Prado-Vázquez G, Chiva C, López-Vacas R, Nanni P, et al. (2017) Prediction of adjuvant chemotherapy response in triple negative breast cancer with discovery and targeted proteomics. PLoS ONE 12 (6): e0178296. https://doi.org/10.1371/journal.pone.0178296

Editor: Aamir Ahmad, University of South Alabama Mitchell Cancer Institute, UNITED STATES

Received: December 19, 2016

Accepted: May 10, 2017

Published: June 8, 2017

Data Availability Statement: All the mass spectrometry raw data files acquired in this study may be downloaded from Chorus (http://chorusproject.org) under the project name Breast Cancer Proteomics. The Parallel Reaction Monitoring dataset is publicly available in the Panorama web server at https://panoramaweb.org/labkey/project/UPF%20-%20CRG/La%20Paz_TN_Breast_Cancer/begin.view?.

Abstract

Background

Triple-negative breast cancer (TNBC) accounts for 15–20% of all breast cancers and usually requires the administration of adjuvant chemotherapy after surgery but even with this treatment many patients still suffer from a relapse. The main objective of this study was to identify proteomics-based biomarkers that predict the response to standard adjuvant chemotherapy, so that patients at are not going to benefit from it can be offered therapeutic alternatives.

Methods

We analyzed the proteome of a retrospective series of formalin-fixed, paraffin-embedded TNBC tissue applying high-throughput label-free quantitative proteomics. We identified several protein signatures with predictive value, which were validated with quantitative targeted proteomics in an independent cohort of patients and further evaluated in publicly available transcriptomics data.

Results

Using univariate Cox analysis, a panel of 18 proteins was significantly associated with distant metastasis-free survival of patients ($p<0.01$). A reduced 5-protein profile with prognostic value was identified and its prediction performance was assessed in an independent targeted proteomics experiment and a publicly available transcriptomics dataset. Predictor *P5*

Funding: We want to particularly acknowledge the patients in this study for their participation and to the IdiPAZ and I+12 Biobanks for the generous gifts of clinical samples used in this work. The IdiPAZ and I+12 Biobanks are supported by Instituto de Salud Carlos III, Spanish Economy and Competitiveness Ministry (RD09/0076/00073 and RD09/0076/00118 respectively) and Farmaindustria, through the Cooperation Program in Clinical and Translational Research of the Community of Madrid. This work was supported by Instituto de Salud Carlos III, Spanish Economy and Competitiveness Ministry, Spain and co-funded by FEDER program, "Una forma de hacer Europa" (PI12/00444, PI12/01016 and PI15/01310). LT-F is supported by Spanish Economy and Competitiveness Ministry (DI-15-07614). The CRG/UPF Proteomics Unit is part of the "Plataforma de Recursos Biomoleculares y Bioinformáticos (ProteoRed)" supported by grant PT13/0001 of ISCIII and Spanish Ministry of Economy and Competitiveness. We acknowledge support of the Spanish Ministry of Economy and Competitiveness, "Centro de Excelencia Severo Ochoa 2013-2017", SEV-2012-0208, and from "Secretaria d'Universitats i Recerca del Departament d'Economia i Coneixement de la Generalitat de Catalunya" (2014SGR678). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing interests: JAFV, AG-P and EE are stakeholders of Biomedica Molecular Medicine S.L. and Biomedica Molecular Medicine Ltd. LT-F is an employee of Biomedica Molecular Medicine S.L. The authors have declared no other conflict of interest. This does not alter our adherence to PLOS ONE policies on sharing data and materials. The authors have declared no other conflict of interest.

Abbreviations: DMFS, distant metastasis-free survival; FDR, false discovery rate; FFPE, formalin-fixed paraffin-embedded; HR, hazard ratio; TNBC, triple negative breast cancer.

including peptides from proteins RAC2, RAB6A, BIEA and IPYR was the best performance protein combination in predicting relapse after adjuvant chemotherapy in TNBC patients.

Conclusions

This study identified a protein combination signature that complements histopathological prognostic factors in TNBC treated with adjuvant chemotherapy. The protein signature can be used in paraffin-embedded samples, and after a prospective validation in independent series, it could be used as predictive clinical test in order to recommend participation in clinical trials or a more exhaustive follow-up.

Introduction

Breast cancer is one of the leading causes of death among women in developed countries. Approximately 20% of the cases correspond to triple-negative tumours, i.e., those not expressing estrogen and progesterone receptors and with no HER2 over-expression. Triple-negative breast cancer (TNBC) is associated with a poor outcome when compared with other subtypes, due to its aggressive behavior and limited therapeutic options [1]. Adjuvant therapy for TNBC relies exclusively on chemotherapy, as hormonal agents and anti-HER2 therapy are no effective in this type of breast cancer. The standard chemotherapy used in this setting includes anthracyclines and taxanes, but even with the use of adjuvant therapy, relapse risk approaches 50% and it is even higher in patients with additional high-risk factors [2].

Moreover, the clinical and molecular heterogeneity within this TNBC subtype makes the treatment of these patients even more challenging as some patients never relapse, whereas others do suffer an early relapse from resistant tumors. Several gene expression profiling evidenced the existence of distinct molecular subgroups of TNBC [3–5]. So far, these molecular studies have not yet allowed the stratification of patients into categories with different prognosis and response to specific treatments. Also, no specific drugs have been developed for the specific treatment of TNBC, although clinical reports suggest a role for platinum compounds [6].

High-throughput technologies for the quantitation of biomolecules are providing a comprehensive view of the molecular changes in cancer tissues. These technologies allow for the simultaneous analysis of the whole genome, global gene and microRNA expression, DNA methylation and protein expression of tumor samples, and in conjunction with the development of bioinformatics tools, have revealed the molecular architecture of breast cancer [7–9]. Recently, two large-scale studies have addressed the structure of the TNBC genome, by means of next generation sequencing and have revealed a plethora of different genetic events occurring in TNBC. Moreover, the results of these studies also revealed the high diversity within this cancer subtype and that there are very few common genetics events in TNBC tumors; mainly a mutation of TP53 that occurs in approximately 80% of these tumors and loss of the tumor suppressor phosphatase PTEN occurring in 29%, with all other mutations occurring at a relatively low frequency [10, 11]. These observations are in agreement with results from other large-scale sequencing studies showing that cancers exhibit extensive mutational heterogeneity, with mutated genes varying widely across individuals [12].

The cellular genotype dictates the observed phenotype through the production of proteins, which, in turn, perform most of the reaction that occur in the cell. Proteomics analyses thus offer a means to measure the biological outcome of cancer-related genomic abnormalities,

including expression of variant proteins encoded by mutations, protein changes driven by altered DNA copy number, chromosomal amplification and deletion events, epigenetic silencing, and changes in microRNA expression [13].

Mass spectrometry has become the method of choice for analyzing complex protein samples, and recent technological advances allow identifying thousands of proteins from tissue amounts compatible with clinical routine. Therefore, proteomics may become a new source of molecular markers with utility in the management of breast cancer patients and to facilitate clinical decisions in daily clinical practice. In the case of TNBC patients, the identification of protein signatures that define patient subgroups that need to be treated with a specific combination of drugs or alternative interventions is highly desirable. In this study, we identified a protein signature with a high prediction value in the response to adjuvant chemotherapy, and validated it in an independent cohort using quantitative targeted proteomics. Indeed, the described protein signature can predict adjuvant chemotherapy response in triple negative breast cancer samples, it is suitable to evaluate formalin-fixed, paraffin-embedded tumour samples, and therefore, it could be used to recommend participation in clinical trials or a more exhaustive follow-up in high-risk TNBC patients.

Materials and methods

Study design and sample description

The discovery cohort comprises twenty-six FFPE samples from patients diagnosed of triple negative breast cancer (TNBC) were retrieved from I+12 Biobank (RD09/0076/00118) and from IdiPAZ Biobank (RD09/0076/00073), both integrated in the Spanish Hospital Biobank Network (RetBioH; www.redbiobancos.es) between 1997 and 2004. The targeted proteomics cohort includes one hundred and fourteen samples from patients diagnosed of triple negative breast cancer were retrieved from I+12 Biobank (RD09/0076/00118) and from IdiPAZ Biobank (RD09/0076/00073), both integrated in the Spanish Hospital Biobank Network (RetBioH; www.redbiobancos.es) between 1997 and 2012. Sixty samples from I+12 Biobank were previously included in an analytical observational case–control study [14]. The histopathological features of each sample were reviewed by an experienced pathologist to confirm diagnosis and tumor content. Eligible samples had to include at least 50% of tumor cells.

Ethics, consent and permissions

Written consent was provided by all patients participating in this study, and approval from the Ethical Committees of Hospitals Doce de Octubre and La Paz was obtained for the conduct of the study.

Total protein extraction

Proteins were extracted from FFPE samples as previously described [15]. Briefly, FFPE sections were deparaffinized in xylene and washed twice with absolute ethanol. Protein extracts from FFPE samples were prepared in 2% SDS buffer using a protocol based on heat-induced antigen retrieval [16]. Protein concentration was determined using the MicroBCA Protein Assay Kit (Pierce-Thermo Scientific). Protein extracts (10 μg) were digested with trypsin (1:50) and SDS was removed from digested lysates using Detergent Removal Spin Columns (Pierce).

Discovery mass spectrometry data acquisition

Samples were analyzed by liquid chromatography-mass spectrometry on a LTQ-Orbitrap Velos (Thermo Fischer Scientific, Bremen, Germany) coupled to NanoLC-Ultra system (Eksigent Technologies, Dublin, CA, USA) as previously described [17]. Peptide samples were further desalted using ZipTips (Millipore), dried, and solubilized in 15 μL of a 0.1% formic acid and 3% acetonitrile solution before MS analysis. Peptide separation was performed on a self-made C18 column (75μm×150mm, 3 μm, 200A) by a 5 to 30% acetonitrile gradient in 95 minutes. Each MS cycle consisted of a full scan MS spectra (300–1700) recorded at resolution of 30000 at 400 m/z followed by CID (collision induced dissociation) fragmentation on the twenty most intense signals. Charge state screening was enabled and singly charge states were rejected. Precursor masses selected for MS/MS were placed in a dynamic exclusion for 45s.

Discovery mass spectrometry data analysis

Protein identification and quantification were performed using the Andromeda search engine and MaxQuant (version 1.2.7.4) [18]. Spectra were searched against a forward UniProtKB/Swiss-Prot database for human concatenated to a reverse decoyed fasta database and containing common protein contaminants. The precursor and fragment tolerances were set respectively to 20ppm and 0.5 Da, carbamidomethyl (C) was set as fixed modification while oxidation (M), deamidation (N, Q) and N-terminal protein acetylation were set as variable modifications. Enzyme specificity was set to Trypsin/P, allowing a minimal peptide length of 7 amino acids and a maximum of two missed cleavages. A maximum false discovery rate (FDR) of 0.01 for peptides and 0.05 for proteins was allowed.

Label free quantification was performed setting a 2 minutes window for match between runs. The protein abundance was calculated on the basis of the normalized spectral protein intensity (LFQ intensity). Quantifiable proteins were defined as those detected in at least 75% of TNBC samples showing two or more unique peptides. Only quantifiable proteins were considered for subsequent analyses. Protein expression data were log2 and missing values were replaced using data imputation for label-free data, as explained in [19], using default values. Finally, protein expression values were z-score transformed. Batch effects were estimated and corrected using ComBat [20].

All the shotgun mass spectrometry raw data files acquired in this study may be downloaded from Chorus (http://chorusproject.org) under the project name *Breast Cancer Proteomics*.

Parallel reaction monitoring data acquisition

Between one and four unique peptides per protein were selected for quantification by parallel reaction monitoring (PRM), prioritizing those peptides that had been observed previously. The selected peptides were bought as isotopically labelled internal standard peptides ($^{13}C_6$,$^{15}N_2$-Lys and $^{13}C_6$,$^{15}N_4$-Arg, Pepotec Peptides, Thermofisher Scientific) and they were spiked in the peptide mixture. The amount spiked-in per for each reference peptide was chosen based on the following criteria: i) to have an area as close to the endogenous peptide area as possible, and ii) to be in within the concentration range in which a linear response of the peptide was observed.

One third of each sample was analyzed using an Orbitrap Fusion Lumos (Thermo Fisher Scientific) coupled to an EASY-nanoLC 1000 UPLC system (Thermo Fisher Scientific) with a 50-cm C18 chromatographic column. Peptide mixes were separated with a chromatographic gradient starting at 5% B with a flow rate of 300 nL/min and going up to 22% B in 79 min and to 32% B in 11 min (Buffer A: 0.1% formic acid in water. Buffer B: 0.1% formic acid in

acetonitrile). The Orbitrap Fusion Lumos was operated in positive ionization mode with an EASY-Spray nanosource at 1.4kV and at a source temperature of 275˚C.

A scheduled PRM method was used for data acquisition with a quadrupole isolation window set to 1.4 m/z and MSMS scans over a mass range of m/z 340–950, with detection in the Orbitrap at a variable resolution depending on the peptide. PRM scans for heavy standards were performed at a resolving power of 15000 (at m/z 200); whereas PRM scans of endogenous peptides were performed at resolution 30000, 60000 or 120000 (at m/z 200) depending on its detectability and observed interferences in previous optimization experiments.

MSMS fragmentation was performed using HCD at 30 NCE, the auto gain control (AGC) was set at 50000 and the injection time (IT) was adjusted according to the transient length, with a maximum of 118 ms for 60000 resolution, and a minimum of 22 ms for 15000 resolution. The size of the scheduled window was 10 min and the maximum cycle time was 2.8 s. All data was acquired with XCalibur software v3.0.63. The Parallel Reaction Monitoring dataset is publicly available in the Panorama web server at https://panoramaweb.org/labkey/project/UPF%20-%20CRG/La%20Paz_TN_Breast_Cancer/begin.view?.

Parallel reaction monitoring data analysis

Product ion chromatographic traces corresponding to the targeted precursor peptides were evaluated with Skyline software v2.5 based on i) traces co-elution, both in its light and heavy forms; and ii) the correlation between the relative intensities of the endogenous product ion traces, and their isotopically-labelled counterparts from the internal reference peptides.

For each monitored peptide a light-to-heavy ratio (L/H ratio = sum of product ion areas of the endogenous peptide/sum product ion areas from the reference peptide) was calculated per patient. Ratios were transformed to the logarithmic scale (log_2) and the obtained values were used as proxy for protein amount.

Prognostic models development and validation

Shotgun data were used to compute a statistical significance level for each protein based on a univariate proportional hazards model [21] with the aim of identifying proteins with an abundance level significantly related to the distant metastasis-free survival (DMFS) as described previously [22]. Briefly, proteins related to DMFS were filtered based on their p-values. Proteins with a p-value$<$0.01 were used to develop prediction models of recurrence risk using the supervised principal component method [23]. Additionally, we evaluated the correlation between the proteins to establish correlation groups and reduce the number of selected proteins to build the molecular signatures. Proteins with a Pearson correlation higher than 0.5 were grouped together and reduced profiles were designed including randomly proteins from different correlation groups. Leave-one-out cross-validation was used to evaluate the predictive accuracy of the profiles. The cutoff point was established *a priori* and to test the statistical significance, the p-value of the log-rank test statistic for the risk groups was evaluated using 1000 random permutations. Analyses were performed in BRB-ArrayTools v4_2_1. BRB-ArrayTools has been developed by Dr. Richard Simon and BRB-ArrayTools Development Team.

Transcriptomics analyses

We used previously published transcriptomics array expression data of 1,296 primary breast carcinomas from two previously published works [24, 25]. Batch effects between data sets were estimated and corrected using ComBat [20]. After protein-to-gene ID conversion, all probes in dataset for each gene were retrieved. Probes with higher coefficient of variation were selected when multiple probes were found for a single gene. We selected estrogen receptor

negative patients with TNBC characteristics, thus we excluded any patient showing an ESR1 relative expression above 12 and ERBB2 relative expression above 11.8, as described previously [26, 27]. Per-gene normalization within the validation cohorts was performed using median values obtained in the discovery cohort. Survival curves were then estimated [28]. Note that no clinical HER2 assessment was available for the transcriptomics samples and that the ERBB2 gene expression value was used for sample classification.

Statistical analyses and software suites

Distant metastasis free survival (DMFS) was defined as the time between the day of surgery and the date of distant relapse or last date of follow-up. The independence of prognostic value of predictors when compared with clinical information was evaluated using multivariate Cox regression analyses. SPSS v16 software package, GraphPad Prism 5.1 and R v2.15.2 (with the *Design* software package 0.2.3) were used for all statistical analyses. All p-values were two-sided and $p<0.05$ was considered statistically signficant.

Results and discussion

Triple-negative breast cancer (TNBC) accounts for one fifth of all breast cancers, and although they are usually treated with the administration of adjuvant chemotherapy after surgery, many patients have a relapse. Therefore, the main objective of this study was to identify proteomics-based biomarkers to stratify patients according to the benefits of the adjuvant chemotherapy, enabling the possibility to offer therapeutic alternatives to patients with predicted poor response to it.

Patient's characteristics

In order to identify prognostic biomarkers of the standard chemotherapy in TNBC patients, we included 25 TNBC patients to be in the discovery study, and 114 TNBC patients to be included in the targeted-proteomics study as an independent validation cohort. The clinical characteristics from all these patients are provided in Table 1. All included patients had node-positive disease; all of the tumors were negative when tested for hormonal receptors using immunohistochemistry and Her2 amplification using immunohistochemistry and fluorescent *in situ* hybridization when needed. Adjuvant chemotherapy was used in all cases (either anthracycline-based or not). In the discovery patient cohort, the median follow-up of all patients was 8.14 years (range: 1.24–12.95) and 9 patients had relapse events. In the validation cohort, median follow-up of all patients was 5.29 years (range: 0.47–11) and 56 patients had relapse events. Adjuvant chemotherapy was used in all patients (either anthracycline-based or not) except in four cases Study design is schematized in Fig 1.

Molecular characterization of TNBC samples by discovery proteomics

Initially, we set up to perform discovery mass spectrometry-based proteomics of the collected 25 FFPE breast cancer samples to identify potential protein candidates that could be used as prognostic biomarkers to chemotherapy response of TNBC patients. Tissue samples were prepared for mass spectrometry analysis with trypsin digestion, following a previously-reported method that exhibit a high reproducibility for these type of samples [23]. Protein abundance data resulting from the mass spectrometry shotgun data acquisition constituted our "*discovery dataset*". One sample was excluded from the study because it was considered an outlier as it did not reach the "mean minus twice the standard deviation"-threshold in the number of unique peptides identified. A total of 3,095 protein groups were identified using the

Table 1. Clinical characteristics of the patients included in the study.

	Discovery cohort	Validation cohort
Age at diagnosis (median)	61.2 (37–78)	57 (25–89)
Age at diagnosis (mean)	58.5	58.9
Tumor Size		
T1	4 (19%)	51 (35.6%)
T2	19 (73%)	109(76.2%)
T3	2 (8%)	7(4.89%)
T4	0 (0%)	8(5.59%)
Multifocal	0 (0%)	1(0.69%)
Tumor Grade		
G1	0 (0%)	4(2.79%)
G2	4 (16%)	22(15.38%)
G3	19 (76%)	112(78.32%)
Unknown	2 (8%)	5(3.49%)
Lymph node status		
N0	0 (0%)	75(52.44%)
N1	17 (68%)	41(28.67%)
N2	8 (32%)	10(6.99%)
N3	0 (0%)	14(9.79%)
Nx	0 (0%)	3(2.09%)
Chemotherapy		
No Antraciclines	11 (42%)	19(16.7%)
Antraciclines	12 (46%)	62(54.3%)
Antraciclines + taxanes	2 (12%)	9(7.9%)
Unknown	0(0%)	20(17.6%)
No	0(0%)	4(3.5%)
Median follow-up (years)	8.14 (1.24–12.95)	5.29 (0.47–11)
Relapse events (%)	9(36%)	56(49%)

Clinical criteria are provided according to TNM classification (http://www.cancer.gov/cancertopics/pdq/treatment/breast/healthprofessional/page3). Tumor grade is the description of a tumor based on how abnormal the tumor cells and the tumor tissue look under a microscope.

https://doi.org/10.1371/journal.pone.0178296.t001

Andromeda database search engine (S1 Table, of which 1,064 presented at least two unique peptides and were detectable in at least 75% of the samples (S2 Table)). Protein label-free quantification was further performed using MaxQuant LFQ values.

In order to identify proteomics-based biomarkers to stratify patients according to the benefits of adjuvant chemotherapy, we performed a survival analysis using the proteins quantified in the discovery dataset and related them with distant metastasis free survival with the Survival Analysis Tool from BRB-ArrayTools. We found that 18 out of 1064 proteins were significantly associated with distant metastasis-free survival (DMFS) of patients in the discovery dataset (Table 2)

Proteomics candidates found in the discovery dataset were also checked in a transcriptomics expression data from 134 triple negative breast cancer samples from two publicly available dataset [24, 25]. To this purpose, per-gene normalization within the validation cohorts was performed. It has been already demonstrated that mRNA levels largely reflect the respective protein levels [29, 30]. Consequently, the intersection between proteomic data sets and other genome-wide data sets often allows robust cross-validation [31, 32].

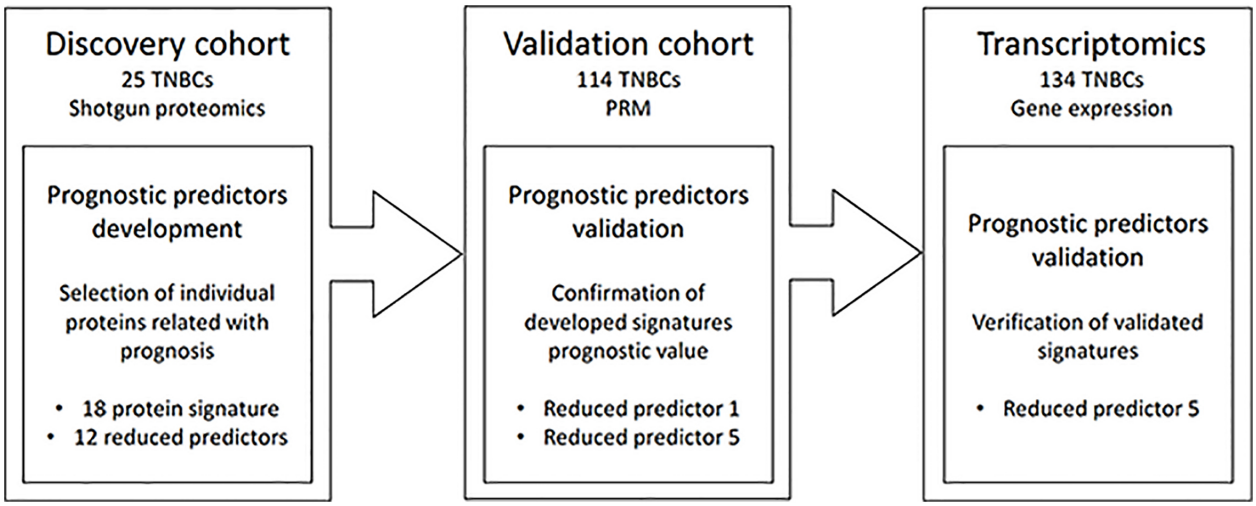


Fig 1. Study design. Chart of samples included and analysis performed in each cohort.

https://doi.org/10.1371/journal.pone.0178296.g001

Identification and validation of prognostic protein based signatures in TNBC patient samples

Protein abundances derived from shotgun mass spectrometry data in the discovery dataset were then used to identify protein combinations with prediction value of distant metastasis free (DMFS) survival after standard chemotherapy. The validation of the prediction value of each proposed protein combination was validated in an independent 114 TNBC patients cohort performing protein quantitation with parallel reaction monitoring approach (PRM), a targeted proteomics approach that enables the quantification of a set of preselected peptides of interest (S3, S4, S5 and S6 Tables). Moreover, proteomics candidates found in the discovery dataset were further assessed in transcriptomics expression data from 134 triple negative breast cancer samples from two publicly available dataset.

Initially, the identified 18 proteins to be significantly associated with DMFS were initially used to build a protein predictor of DMFS containing all 18 proteins. The cutoff threshold value was bounded *a priori* to split the population with a 50:50 distribution between low and high distant metastasis risk. DMFS at 5 years was 100% for patients defined as low-risk by the prognostic profile versus 25% for patients defined as high-risk (hazard ratio (HR) = 16.36, $p<0.0001$). However, the prognostic value of this signature could not be validated neither using PRM data from the validation cohort nor using the publicly available transcriptomics dataset. In the PRM validation cohort, DMFS at 5 years was 59.8% for patients defined as low-risk by the prognostic profile versus 56.6% for patients defined as high-risk when used a 50:50 cutoff value (HR = 1.065, p = 0.78). In the transcriptomics verification, when using a 50:50 cut-off, DMFS at 5 years was 71.3% for patients defined as low-risk by the prognostic profile versus 66.5% for patients defined as high-risk (HR = 1.309, p = 0.38).

We then explored the possibility of developing a protein combination using a reduced number of proteins, as the incorporation of redundant information may reduce the chances of finding a valid predictor [28]. Towards this direction, we established three groups of proteins based on the correlation of their expression abundance patterns and one or two proteins belonging to different correlation groups were randomly included to build predictors that included three to seven proteins. Again, a 50:50 distribution between low and high distant metastasis risk was set *a priori* to obtain a cutoff threshold value. Twelve protein combinations

Table 2. Proteins significantly associated with distant metastasis free survival.

UniProtKB accession numbers	Uniprot ID	Protein name	Gene Symbol	Hazard ratio	P value
O43175	SERA_HUMAN	D-3-phosphoglycerate dehydrogenase (3-PGDH) (EC 1.1.1.95)	PHGDH PGDH3	0.689	0.001
O75323	NIPS2_HUMAN	Protein NipSnap homolog 2 (NipSnap2) (Glioblastoma-amplified sequence)	GBAS NIPSNAP2	1.830	0.001
P05091	ALDH2_HUMAN	Aldehyde dehydrogenase, mitochondrial (EC 1.2.1.3) (ALDH class 2) (ALDHI)	ALDH2 ALDM	0.423	0.002
P05161	ISG15_HUMAN	Ubiquitin-like protein ISG15 (Interferon-induced 15 kDa protein) (Interferon-induced 17 kDa protein) (IP17) (Ubiquitin cross-reactive protein) (hUCRP)	ISG15 G1P2 UCRP	0.500	0.002
P07996	TSP1_HUMAN	Thrombospondin-1	THBS1 TSP TSP1	0.649	0.002
P14317	HCLS1_HUMAN	Hematopoietic lineage cell-specific protein (Hematopoietic cell-specific LYN substrate 1) (LckBP1) (p75)	HCLS1 HS1	0.379	0.003
P15153	RAC2_HUMAN	Ras-related C3 botulinum toxin substrate 2 (GX) (Small G protein) (p21-Rac2)	RAC2	0.423	0.003
P18085	ARF4_HUMAN	ADP-ribosylation factor 4	ARF4 ARF2	3.754	0.003
P20340	RAB6A_HUMAN	Ras-related protein Rab-6A (Rab-6)	RAB6A RAB6	0.493	0.004
P28065	PSB9_HUMAN	Proteasome subunit beta type-9 (EC 3.4.25.1) (Low molecular mass protein 2) (Proteasome subunit beta-1i) (Really interesting new gene 12 protein)	PSMB9 LMP2 PSMB6i RING12	0.758	0.005
P53004	BIEA_HUMAN	Biliverdin reductase A (BVR A) (EC 1.3.1.24) (Biliverdin-IX alpha-reductase)	BLVRA BLVR BVR	0.674	0.006
P62873	GBB1_HUMAN	Guanine nucleotide-binding protein G(I)/G(S)/G(T) subunit beta-1 (Transducin beta chain 1)	GNB1	0.703	0.006
Q09666	AHNK_HUMAN	Neuroblast differentiation-associated protein AHNAK (Desmoyokin)	AHNAK PM227	1.614	0.006
Q15046	SYK_HUMAN	Lysine—tRNA ligase (EC 6.1.1.6) (Lysyl-tRNA synthetase) (LysRS)	KARS KIAA0070	0.672	0.008
Q15181	IPYR_HUMAN	Inorganic pyrophosphatase (EC 3.6.1.1) (Pyrophosphate phospho-hydrolase)	PPA1 IOPPP PP	2.184	0.008
Q9BUP0	EFHD1_HUMAN	EF-hand domain-containing protein D1 (EF-hand domain-containing protein 1) (Swiprosin-2)	EFHD1 SWS2 PP3051	0.265	0.009
Q9GZZ9	UBA5_HUMAN	Ubiquitin-like modifier-activating enzyme 5 (Ubiquitin-activating enzyme 5) (ThiFP1) (UFM1-activating enzyme) (Ubiquitin-activating enzyme E1 domain-containing protein 1)	UBA5 UBE1DC1	0.316	0.009
Q9NR31	SAR1A_HUMAN	GTP-binding protein SAR1a (COPII-associated small GTPase)	SAR1A SAR1 SARA SARA1	0.222	0.009

These 18 proteins are significant with $p < 0.01$ in the univariate test.

https://doi.org/10.1371/journal.pone.0178296.t002

were built and they all exhibited a significant prognostic value in our discovery dataset (S1 Fig and S7 Table).

Using the protein abundances derived from the PRM analysis of the 114 TNBC tumor samples, we could validate two out of twelve reduced predictors, which also showed a significant prognostic value in an independent cohort of patients (Table 3). Predictor P1 showed a significant prognostic value using a 70:30 distribution between low and high risk patients. DMFS at 5-years was of 65.6% in the low-risk group and 29.92% at high-risk group (HR = 2.577, p = 0.0002). Predictor *P5* showed a significant prognostic value using a 70:30 distribution between low and high risk patients. DMFS at 5-years was of 63.54% in the low-risk group and 39.99% at high-risk group (HR = 2.322, p = 0.0142). Moreover, predictor *P5* also showed a significant prognostic value when compared with tumor size and lymph node status using multivariate Cox regression analyses (S8 and S9 Tables), and when used to predict the behavior of the patients analyzed in the transcriptomics dataset.

Table 3. DMFS prediction of the two reduced predictors tested in the publicly available transcriptomics dataset.

Reducedpredictor	ProteinID	DMFS[$](low risk)	DMFS[$](high risk)	HR(95%CI)	p	DMFS[$](low risk)	DMFS[$](high risk)	HR (95%CI)	p
		70:30 cutoff				50:50 cutoff			
Predictor_1PRM validation	P53004P05161P28065O75323	67.1%	29.9%	3.277(1.740–6.172)	>0.01	73.9%	37.2%	3.094(1.906–5.540)	>0.01
Predictor_5PRM validation	P53004P20340P15153Q15181	63.5%	39.9%	1.774(1.057.–3.453)	0.01	61.7%	48.8%	1.327(0.787–2.246)	0.15
		Defined cutoff				50:50 cutoff			
Predictor_1Transcriptomics	P53004P05161P28065O75323	72.0%	68. %	1.311(0.647–2.668)	0.45	67.7%	71.0%	0.907(0.495–1.66)	0.75
Predictor_5Transcriptomics	P53004P20340P15153Q15181	80.9%	66.6%	1.837(0.844–3.998)	0.13	76.4%	63. %	1.888(1.027–3.468)	0.041

[$]DMFS is calculated at five years

https://doi.org/10.1371/journal.pone.0178296.t003

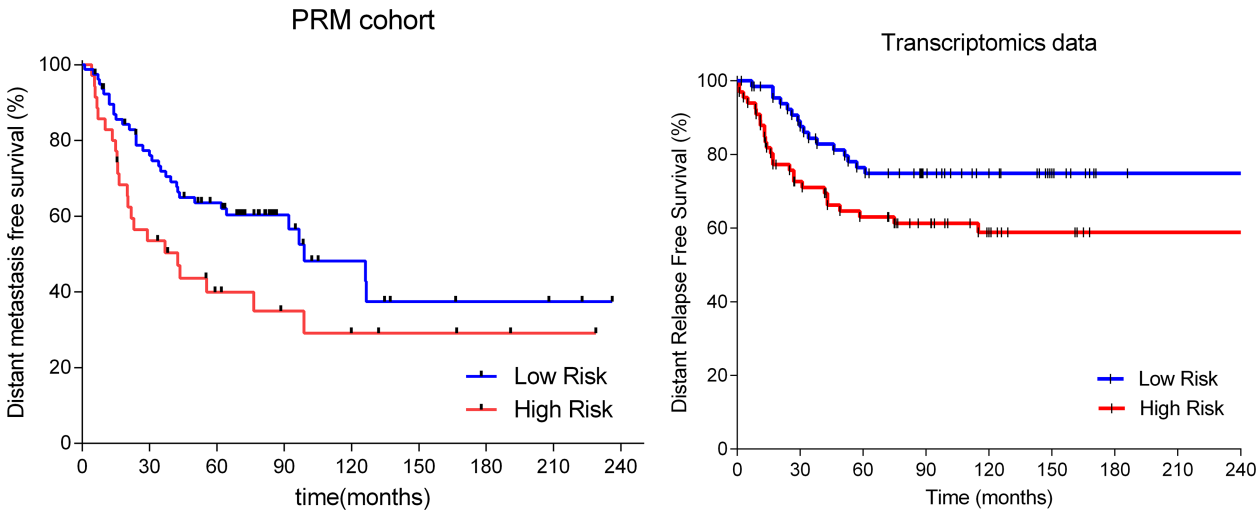


Fig 2. Survival analysis of reduced profile 5 in the PRM validation cohort and in the trasncriptomics orthogonal verification.

https://doi.org/10.1371/journal.pone.0178296.g002

Finally, we also checked the performance of the reduced predictors *P1* and *P5* in the two publicly available transcriptomics datasets. In these data, predictor *P1* showed no prognostic information, whereas predictor *P5* showed a DMFS in the low-risk group over 80% using the test set defined cutoff thresholds, but they assigned less than 20% of the patients to this group. However, this last results leaves too many patients who do not relapse in the high-risk group, and thus, we tested a 50:50 cutoff threshold in this predictor. When a 50:50 cutoff threshold was used DMFS at five years in the publicly available transcriptomics dataset was 78.0% for low-risk patients versus 61.4% (HR = 2.888, p = 0.041) (Table 3 and Fig 2).

Predictor *P5* includes peptides from proteins RAC2, RAB6A, BIEA and IPYR. RAC2 is a member of the Ras superfamily of small guanosine triphosphate (GTP)-metabolizing proteins. It has been proposed that protein RAC2 might have a role in the regulation of the actin cytoskeleton during breast cancer metastasis [33]. RAC2 is also involved in both PLD-induced cell invasion [34] and oncogenic KIT-induced neoplasms [35], and its under-expression has been related to invasive and metastatic competence in human cancer [36]. BIEA, the protein encoded by the biliverdin reductase A (BLVRA) gene, belongs to the biliverdin reductase family members, which catalyze the conversion of biliverdin to bilirubin in the presence of NADPH or NADH. It also works as a dual-specificity kinase (S/T/Y), and activates the MAPK and IGF/IRK receptor signal transduction pathways [37, 38]. BIEA plays a pivotal role in the development of multidrug resistance in human HL60 leukemia cells [39], and itis included among the 50 genes that compose the PAM50 gene signature for classifying "intrinsic" subtypes of breast cancer [40].

RAB6A is a member of the RAB family, which belongs to the small GTPase superfamily. This protein is located at the Golgi apparatus, which regulates protein-trafficking. RAB6A is a potential target of both miR-21 and miR-155, known to be deregulated [41] and be correlated with a poor prognosis in breast cancer [42–44], which supports our findings. Additionally, RAB6A showed an increased expression in the HER-2/neu breast cancer subgroup [45].

Finally, IPYR is a cytosolic inorganic pyrophosphatase, codified by the PPA1 gene. PPA1 expression is significantly higher in many tumors, especially those of lung and ovarian origin. Expression of IPYR is heterogeneous in breast cancer cells [46] and the knockdown of PPA1 shows a decreased colony formation and viability of MCF7 cells [47]. Additionally,

pyrophosphatase overexpression has been associated with cell migration, invasion, and poor prognosis in gastric cancer [48].

Conclusions

High-throughput proteomics can be used to identify subgroups with different prognosis among patients with TNBC and to derive signatures with a combination of multiple proteins that enable patient stratification. Defining multi-gene or multi-protein predictors for prognosis increases their accuracy, reproducibility and robustness, which are highly desirable features in clinical diagnostic and prognostic tools. Towards this direction, Liu and colleagues developed a 11-protein signature in early triple-negative breast cancer [49] which showed a prognostic value in lymph node negative patient who had not received systemic adjuvant therapy. The protein signature was validated in an independent dataset using a cutoff determined from the ROC curve of the training set to ensure high-sensitivity and specificity. However, for validation purposes it is usually important that cutoff thresholds of a risk score be defined in advance [50]. Other authors have defined prognostic and predictive signatures in TNBCs using gene expression measurement techniques [4, 51, 52].

In the present work, we described the first protein-based signatures to predict adjuvant chemotherapy response in triple negative breast cancer samples. Several protein predictors were derived from a shotgun mass spectrometry-based discovery dataset and their performance was further validated in an independent patient cohort using targeted proteomics (parallel reaction monitoring). Our protein signatures were derived from routinely processed FFPE samples on a population of TNBC patients treated with adjuvant chemotherapy, which is closer to the clinical reality. Within these context, predictor *P5* that includes peptides from proteins RAC2, RAB6A, BIEA and IPYR, emerged as the best predictor when accounting both the discovery and the validation proteomics datasets. Moreover, its performance was also confirmed in a publicly available transcriptomics dataset, which exemplify the robustness of the described predictor and its applicability to patient-derived transcriptomics data that might be already collected.

Although our findings require prospective validation in independent series for routine clinical application, our work demonstrates the potential of proteomics to assist oncologists to make clinical decisions regarding patient treatment; e.g., patients classified with the low-risk group by the identified protein signature need to be treated with standard chemotherapy, whereas those classified with the high-risk group should be offered clinical trials with new drugs and an intensive follow-up program.

Supporting information

S1 Fig. Kaplan-Meier graphs of reduced profiles.
(TIF)

S1 Table. Shotgun proteomics LFQ values.
(TXT)

S2 Table. Log_2 transformed and normalized protein expression data.
(TXT)

S3 Table. Sample and patient codes of PRM analyses.
(XLSX)

S4 Table. Scheduled PRM Method for Orbitrap Fusion Lumos.
(XLSX)

S5 Table. Product ion area for quantified endogenous and isotopically-labelled peptides.
(XLSX)

S6 Table. Log2 ratio of the areas of the quantified endogenous and isotopically-labelled peptides.
(XLSX)

S7 Table. Survival analysis of reduced profiles in the discovery cohort.
(DOCX)

S8 Table. Multivariate Cox regression model in discovery cohort. T: tumor size, N: lymph node status, HR: Hazard Ratio.
(DOCX)

S9 Table. Multivariate Cox regression model in targeted-proteomics cohort. T: tumor size, N: lymph node status, HR: Hazard Ratio.
(DOCX)

Acknowledgments

We want to particularly acknowledge the patients in this study for their participation and to the IdiPAZ and I+12 Biobanks for the generous gifts of clinical samples used in this work. The IdiPAZ and I+12 Biobanks are supported by Instituto de Salud Carlos III, Spanish Economy and Competitiveness Ministry (RD09/0076/00073 and RD09/0076/00118 respectively) and Farmaindustria, through the Cooperation Program in Clinical and Translational Research of the Community of Madrid. This work was supported by Instituto de Salud Carlos III, Spanish Economy and Competitiveness Ministry, Spain and co-funded by FEDER program, "Una forma de hacer Europa" (PI12/00444, PI12/01016 and PI15/01310). LT-F is supported by Spanish Economy and Competitiveness Ministry (DI-15-07614). The CRG/UPF Proteomics Unit is part of the "Plataforma de Recursos Biomoleculares y Bioinformáticos (ProteoRed)" supported by grant PT13/0001 of ISCIII and Spanish Ministry of Economy and Competitiveness. We acknowledge support of the Spanish Ministry of Economy and Competitiveness, "Centro de Excelencia Severo Ochoa 2013–2017", SEV-2012-0208, and from "Secretaria d'Universitats i Recerca del Departament d'Economia i Coneixement de la Generalitat de Catalunya" (2014SGR678). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Author Contributions

Conceptualization: AG-P EC EE JAFV.

Data curation: JG AG-P JB-S GP-V LT-F.

Formal analysis: MD-A JG.

Funding acquisition: EC JAFV.

Investigation: AG-P LT-F JB-S GP-V RL-V PN ES CC.

Resources: EC EE.

Supervision: JAFV.

Validation: JAFV AGP MD-A.

Writing – original draft: AG-P JAF-V.

Writing – review & editing: JAF-V LT-F AG-P.

References

1. Dent R, Trudeau M, Pritchard KI, Hanna WM, Kahn HK, Sawka CA, et al. Triple-negative breast cancer: clinical features and patterns of recurrence. Clin Cancer Res. 2007; 13(15 Pt 1):4429–34. Epub 2007/08/03. https://doi.org/10.1158/1078-0432.CCR-06-3045 PMID: 17671126.
2. Albergaria A, Ricardo S, Milanezi F, Carneiro V, Amendoeira I, Vieira D, et al. Nottingham Prognostic Index in triple-negative breast cancer: a reliable prognostic tool? BMC Cancer. 2011; 11:299. Epub 2011/07/15. https://doi.org/10.1186/1471-2407-11-299 PMID: 21762477;
3. Hirshfield KM, Ganesan S. Triple-negative breast cancer: molecular subtypes and targeted therapy. Curr Opin Obstet Gynecol. 2014; 26(1):34–40. https://doi.org/10.1097/GCO.0000000000000038 PMID: 24346128.
4. Lee U, Frankenberger C, Yun J, Bevilacqua E, Caldas C, Chin SF, et al. A prognostic gene signature for metastasis-free survival of triple negative breast cancer patients. PLoS One. 2013; 8(12):e82125. Epub 2013/12/11. https://doi.org/10.1371/journal.pone.0082125 PMID: 24349199;
5. Lehmann BD, Bauer JA, Chen X, Sanders ME, Chakravarthy AB, Shyr Y, et al. Identification of human triple-negative breast cancer subtypes and preclinical models for selection of targeted therapies. J Clin Invest. 2011; 121(7):2750–67. https://doi.org/10.1172/JCI45014 PMID: 21633166;
6. André F, Zielinski CC. Optimal strategies for the treatment of metastatic triple-negative breast cancer with currently approved agents. Ann Oncol. 2012; 23 Suppl 6:vi46-51. https://doi.org/10.1093/annonc/mds195 PMID: 23012302.
7. Aebersold R, Mann M. Mass spectrometry-based proteomics. Nature. 2003; 422(6928):198–207. https://doi.org/10.1038/nature01511 PMID: 12634793.
8. Aebersold R, Mann M. Mass-spectrometric exploration of proteome structure and function. Nature. 2016; 537(7620):347–55. https://doi.org/10.1038/nature19949 PMID: 27629641.
9. Mertins P, Mani DR, Ruggles KV, Gillette MA, Clauser KR, Wang P, et al. Proteogenomics connects somatic mutations to signalling in breast cancer. Nature. 2016; 534(7605):55–62. Epub 2016/05/25. https://doi.org/10.1038/nature18003 PMID: 27251275;
10. Network CGA. Comprehensive molecular portraits of human breast tumours. Nature. 2012; 490(7418):61–70. Epub 2012/09/23. https://doi.org/10.1038/nature11412 PMID: 23000897;
11. Shah SP, Roth A, Goya R, Oloumi A, Ha G, Zhao Y, et al. The clonal and mutational evolution spectrum of primary triple-negative breast cancers. Nature. 2012; 486(7403):395–9. Epub 2012/04/04. https://doi.org/10.1038/nature10933 PMID: 22495314;
12. Raphael BJ. Making connections: using networks to stratify human tumors. Nat Methods. 2013; 10(11):1077–8. https://doi.org/10.1038/nmeth.2704 PMID: 24173383.
13. Ellis MJ, Gillette M, Carr SA, Paulovich AG, Smith RD, Rodland KK, et al. Connecting genomic alterations to cancer biology with proteomics: the NCI Clinical Proteomic Tumor Analysis Consortium. Cancer Discov. 2013; 3(10):1108–12. https://doi.org/10.1158/2159-8290.CD-13-0219 PMID: 24124232;
14. Laimito KR, Gámez-Pozo A, Sepúlveda J, Manso L, López-Vacas R, Pascual T, et al. Characterisation of the triple negative breast cancer phenotype associated with the development of central nervous system metastases. Ecancermedicalscience. 2016; 10:632. Epub 2016/04/11. https://doi.org/10.3332/ecancer.2016.632 PMID: 27170832;
15. Gamez-Pozo A, Ferrer NI, Ciruelos E, Lopez-Vacas R, Martinez FG, Espinosa E, et al. Shotgun proteomics of archival triple-negative breast cancer samples. Proteomics Clin Appl. 2013; 7(3–4):283–91. Epub 2013/02/26. https://doi.org/10.1002/prca.201200048 PMID: 23436753.
16. Gámez-Pozo A, Sánchez-Navarro I, Calvo E, Díaz E, Miguel-Martín M, López R, et al. Protein phosphorylation analysis in archival clinical cancer samples by shotgun and targeted proteomics approaches. Mol Biosyst. 2011; 7(8):2368–74. https://doi.org/10.1039/c1mb05113j PMID: 21617801.
17. Gámez-Pozo A, Berges-Soria J, Arevalillo JM, Nanni P, López-Vacas R, Navarro H, et al. Combined label-free quantitative proteomics and microRNA expression analysis of breast cancer unravel molecular differences with clinical implications. Cancer Res; 2015. p. 2243–53. https://doi.org/10.1158/0008-5472.CAN-14-1937 PMID: 25883093
18. Cox J, Mann M. MaxQuant enables high peptide identification rates, individualized p.p.b.-range mass accuracies and proteome-wide protein quantification. Nat Biotechnol. 2008; 26(12):1367–72. Epub 2008/11/26. https://doi.org/10.1038/nbt.1511 PMID: 19029910.
19. Deeb SJ, D'Souza RC, Cox J, Schmidt-Supprian M, Mann M. Super-SILAC allows classification of diffuse large B-cell lymphoma subtypes by their protein expression profiles. Mol Cell Proteomics. 2012; 11(5):77–89. Epub 2012/03/21. https://doi.org/10.1074/mcp.M111.015362 PMID: 22442255;

20. Johnson WE, Li C, Rabinovic A. Adjusting batch effects in microarray expression data using empirical Bayes methods. Biostatistics. 2007; 8(1):118–27. https://doi.org/10.1093/biostatistics/kxj037 PMID: 16632515.
21. Cox DR. Regression models and life-tables. *J Roy Stat Soc*.1972. p. 187–220.
22. Ouyang M, Li Y, Ye S, Ma J, Lu L, Lv W, et al. MicroRNA profiling implies new markers of chemoresistance of triple-negative breast cancer. PLoS One. 2014; 9(5):e96228. Epub 2014/05/03. https://doi.org/10.1371/journal.pone.0096228 PMID: 24788655;
23. Bair E, Tibshirani R. Semi-supervised methods to predict patient survival from gene expression data. PLoS Biol. 2004; 2(4):E108. Epub 2004/04/13. https://doi.org/10.1371/journal.pbio.0020108 PMID: 15094809;
24. Guedj M, Marisa L, de Reynies A, Orsetti B, Schiappa R, Bibeau F, et al. A refined molecular taxonomy of breast cancer. Oncogene. 2012; 31(9):1196–206. https://doi.org/10.1038/onc.2011.301 PMID: 21785460;
25. Miller LD, Coffman LG, Chou JW, Black MA, Bergh J, D'Agostino R, et al. An iron regulatory gene signature predicts outcome in breast cancer. Cancer Res. 2011; 71(21):6728–37. https://doi.org/10.1158/0008-5472.CAN-11-1870 PMID: 21875943;
26. Bianchini G, Iwamoto T, Qi Y, Coutant C, Shiang CY, Wang B, et al. Prognostic and therapeutic implications of distinct kinase expression patterns in different subtypes of breast cancer. Cancer Res. 2010; 70 (21):8852–62. Epub 2010/10/19. https://doi.org/10.1158/0008-5472.CAN-10-1039 PMID: 20959472.
27. Gong Y, Yan K, Lin F, Anderson K, Sotiriou C, Andre F, et al. Determination of oestrogen-receptor status and ERBB2 status of breast carcinoma: a gene-expression profiling study. Lancet Oncol. 2007; 8 (3):203–11. https://doi.org/10.1016/S1470-2045(07)70042-6 PMID: 17329190.
28. Sánchez-Navarro I, Gámez-Pozo A, Pinto A, Hardisson D, Madero R, López R, et al. An 8-gene qRT-PCR-based gene expression score that has prognostic value in early breast cancer. BMC Cancer. 2010; 10:336. Epub 2010/06/28. https://doi.org/10.1186/1471-2407-10-336 PMID: 20584321;
29. Marguerat S, Schmidt A, Codlin S, Chen W, Aebersold R, Bähler J. Quantitative analysis of fission yeast transcriptomes and proteomes in proliferating and quiescent cells. Cell. 2012; 151(3):671–83. https://doi.org/10.1016/j.cell.2012.09.019 PMID: 23101633;
30. Nagaraj N, Wisniewski JR, Geiger T, Cox J, Kircher M, Kelso J, et al. Deep proteome and transcriptome mapping of a human cancer cell line. Mol Syst Biol. 2011; 7:548. Epub 2011/11/08. https://doi.org/10.1038/msb.2011.81 PMID: 22068331;
31. Tyers M, Mann M. From genomics to proteomics. Nature. 2003; 422(6928):193–7. https://doi.org/10.1038/nature01510 PMID: 12634792.
32. Liu Y, Beyer A, Aebersold R. On the Dependency of Cellular Protein Levels on mRNA Abundance. Cell. 2016; 165(3):535–50. https://doi.org/10.1016/j.cell.2016.03.014 PMID: 27104977.
33. Li H, Yang L, Fu H, Yan J, Wang Y, Guo H, et al. Association between Galphai2 and ELMO1/Dock180 connects chemokine signalling with Rac activation and metastasis. Nat Commun. 2013; 4:1706. Epub 2013/04/18. https://doi.org/10.1038/ncomms2680 PMID: 23591873;
34. Henkels KM, Boivin GP, Dudley ES, Berberich SJ, Gomez-Cambronero J. Phospholipase D (PLD) drives cell invasion, tumor growth and metastasis in a human breast cancer xenograph model. Oncogene. 2013; 32(49):5551–62. Epub 2013/06/10. https://doi.org/10.1038/onc.2013.207 PMID: 23752189;
35. Martin H, Mali RS, Ma P, Chatterjee A, Ramdas B, Sims E, et al. Pak and Rac GTPases promote oncogenic KIT-induced neoplasms. J Clin Invest. 2013; 123(10):4449–63. Epub 2013/09/16. https://doi.org/10.1172/JCI67509 PMID: 24091327;
36. Gildea JJ, Seraj MJ, Oxford G, Harding MA, Hampton GM, Moskaluk CA, et al. RhoGDI2 is an invasion and metastasis suppressor gene in human cancer. Cancer Res. 2002; 62(22):6418–23. Epub 2002/11/20. PMID: 12438227.
37. Gibbs PE, Maines MD. Biliverdin inhibits activation of NF-kappaB: reversal of inhibition by human biliverdin reductase. Int J Cancer. 2007; 121(11):2567–74. https://doi.org/10.1002/ijc.22978 PMID: 17683071.
38. Lerner-Marmarosh N, Miralem T, Gibbs PE, Maines MD. Human biliverdin reductase is an ERK activator; hBVR is an ERK nuclear transporter and is required for MAPK signaling. Proc Natl Acad Sci U S A. 2008; 105(19):6870–5. Epub 2008/05/07. https://doi.org/10.1073/pnas.0800750105 PMID: 18463290;
39. Kim SS, Seong S, Lim SH, Kim SY. Targeting biliverdin reductase overcomes multidrug resistance in leukemia HL60 cells. Anticancer Res. 2013; 33(11):4913–9. PMID: 24222129.
40. Parker JS, Mullins M, Cheang MC, Leung S, Voduc D, Vickery T, et al. Supervised risk predictor of breast cancer based on intrinsic subtypes. J Clin Oncol. 2009; 27(8):1160–7. Epub 2009/02/09. https://doi.org/10.1200/JCO.2008.18.1370 PMID: 19204204;

41. Iorio MV, Ferracin M, Liu CG, Veronese A, Spizzo R, Sabbioni S, et al. MicroRNA gene expression deregulation in human breast cancer. Cancer Res. 2005; 65(16):7065–70. Epub 2005/08/17. https://doi.org/10.1158/0008-5472.CAN-05-1783 PMID: 16103053.

42. Chen J, Wang BC, Tang JH. Clinical significance of microRNA-155 expression in human breast cancer. J Surg Oncol. 2012; 106(3):260–6. Epub 2011/11/21. https://doi.org/10.1002/jso.22153 PMID: 22105810.

43. Lee JA, Lee HY, Lee ES, Kim I, Bae JW. Prognostic Implications of MicroRNA-21 Overexpression in Invasive Ductal Carcinomas of the Breast. J Breast Cancer. 2011; 14(4):269–75. Epub 2011/12/27. https://doi.org/10.4048/jbc.2011.14.4.269 PMID: 22323912;

44. Yan LX, Huang XF, Shao Q, Huang MY, Deng L, Wu QL, et al. MicroRNA miR-21 overexpression in human breast cancer is associated with advanced clinical stage, lymph node metastasis and patient poor prognosis. RNA. 2008; 14(11):2348–60. Epub 2008/09/23. https://doi.org/10.1261/rna.1034808 PMID: 18812439;

45. Sotiriou C, Neo SY, McShane LM, Korn EL, Long PM, Jazaeri A, et al. Breast cancer classification and prognosis based on gene expression profiles from a population-based study. Proc Natl Acad Sci U S A. 2003; 100(18):10393–8. Epub 2003/08/13. https://doi.org/10.1073/pnas.1732912100 PMID: 12917485;

46. Luo D, Wang G, Shen W, Zhao S, Zhou W, Wan L, et al. Clinical significance and functional validation of PPA1 in various tumors. Cancer Med. 2016; 5(10):2800–12. Epub 2016/09/26. https://doi.org/10.1002/cam4.894 PMID: 27666431;

47. Mishra DR, Chaudhary S, Krishna BM, Mishra SK. Identification of Critical Elements for Regulation of Inorganic Pyrophosphatase (PPA1) in MCF7 Breast Cancer Cells. PLoS One. 2015; 10(4):e0124864. Epub 2015/04/29. https://doi.org/10.1371/journal.pone.0124864 PMID: 25923237;

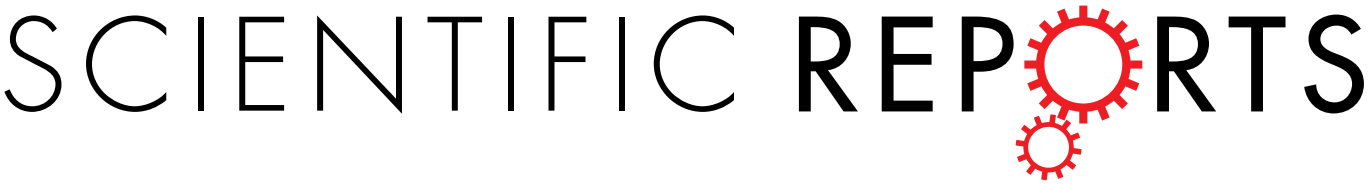
48. Jeong SH, Ko GH, Cho YH, Lee YJ, Cho BI, Ha WS, et al. Pyrophosphatase overexpression is associated with cell migration, invasion, and poor prognosis in gastric cancer. Tumour Biol. 2012; 33(6):1889–98. Epub 2012/07/14. https://doi.org/10.1007/s13277-012-0449-5 PMID: 22797819.

49. Liu NQ, Stingl C, Look MP, Smid M, Braakman RB, De Marchi T, et al. Comparative proteome analysis revealing an 11-protein signature for aggressive triple-negative breast cancer. J Natl Cancer Inst. 2014; 106(2):djt376. Epub 2014/01/07. https://doi.org/10.1093/jnci/djt376 PMID: 24399849;

50. Simon R. Roadmap for developing and validating therapeutically relevant genomic classifiers. J Clin Oncol. 2005; 23(29):7332–41. Epub 2005/09/06. https://doi.org/10.1200/JCO.2005.02.8712 PMID: 16145063.

51. Yau C, Esserman L, Moore DH, Waldman F, Sninsky J, Benz CC. A multigene predictor of metastatic outcome in early stage hormone receptor-negative and triple-negative breast cancer. Breast Cancer Res. 2010; 12(5):R85. Epub 2010/10/14. https://doi.org/10.1186/bcr2753 PMID: 20946665;

52. Yu KD, Zhu R, Zhan M, Rodriguez AA, Yang W, Wong S, et al. Identification of prognosis-relevant subgroups in patients with chemoresistant triple-negative breast cancer. Clin Cancer Res. 2013; 19(10):2723–33. Epub 2013/04/02. https://doi.org/10.1158/1078-0432.CCR-12-2986 PMID: 23549873



OPEN

Urothelial cancer proteomics provides both prognostic and functional information

Received: 14 July 2017
Accepted: 1 November 2017
Published online: 17 November 2017

Guillermo de Velasco[1], Lucia Trilla-Fuertes[2,3], Angelo Gamez-Pozo[2,3], Maria Urbanowicz[4], Gustavo Ruiz-Ares[1], Juan M. Sepúlveda[1], Guillermo Prado-Vazquez[2], Jorge M. Arevalillo[5], Andrea Zapater-Moros[2], Hilario Navarro[5], Rocio Lopez-Vacas[2], Ray Manneh[1], Irene Otero[1], Felipe Villacampa[6,8], Jesus M. Paramio[7,8], Juan Angel Fresno Vara[2,3,8] & Daniel Castellano[1,8]

Traditionally, bladder cancer has been classified based on histology features. Recently, some works have proposed a molecular classification of invasive bladder tumors. To determine whether proteomics can define molecular subtypes of muscle invasive urothelial cancer (MIUC) and allow evaluating the status of biological processes and its clinical value. 58 MIUC patients who underwent curative surgical resection at our institution between 2006 and 2012 were included. Proteome was evaluated by high-throughput proteomics in routinely archive FFPE tumor tissue. New molecular subgroups were defined. Functional structure and individual proteins prognostic value were evaluated and correlated with clinicopathologic parameters. 1,453 proteins were quantified, leading to two MIUC molecular subgroups. A protein-based functional structure was defined, including several nodes with specific biological activity. The functional structure showed differences between subtypes in metabolism, focal adhesion, RNA and splicing nodes. Focal adhesion node has prognostic value in the whole population. A 6-protein prognostic signature, associated with higher risk of relapse (5 year DFS 70% versus 20%) was defined. Additionally, we identified two MIUC subtypes groups. Prognostic information provided by pathologic characteristics is not enough to understand MIUC behavior. Proteomics analysis may enhance our understanding of prognostic and classification. These findings can lead to improving diagnosis and treatment selection in these patients.

Urothelial cancer (UC) is responsible for approximately 165,000 deaths per year worldwide (GLOBOCAN 2012)[1]. Pathological classification divides UC into two major subtypes according to the invasion depth: non-muscle invasive and muscle invasive urothelial carcinoma (MIUC) but not molecular categorization is clinically indicated. However, the outcome and prognosis may be different across subsets of patients within same staging.

MIUC is characterized by a high risk of relapse and metastasise. Despite radical cystectomy with neoadjuvant cisplatin-based chemotherapy, the current risk of recurrence as well as mortality is nearly 50%[2]. In the adjuvant setting, chemotherapy is also associated with improved survival in patients with locally advanced bladder cancer[3].

Pathological prognostic factors such as lymphovascular invasion, grade or molecular alterations are not currently modifying treatment choice. Large collaborative efforts have provided a more comprehensive view of the genomic landscape of MIUC identifying molecular subtypes that have yet to prove predictive value[3–5]. At present, no molecularly targeted drugs are approved for UC.

Before the genomic era, p53 was thought to be prognostic and predictive marker measured by immunohistochemistry in UC[6]. Several methodological issues questioned conflicting results including proteomics assessment[7].

[1]Department of Medical Oncology, University Hospital 12 de Octubre, i+12, Madrid, Spain. [2]Molecular Oncology & Pathology Lab, INGEMM, Instituto de Investigación Hospital La Paz-IdiPAZ, Madrid, Spain. [3]Biomedica Molecular Medicine, Madrid, Spain. [4]Department of Pathology, University Hospital 12 de Octubre, Madrid, Spain. [5]Department of Statistics, Operational Research and Numerical Analysis, University Nacional Educacion a Distancia (UNED), Madrid, Spain. [6]Department of Urology, University Hospital 12 de Octubre, Madrid, Spain. [7]Molecular and Cell Oncology Group, Biomedical research Institute, University Hospital 12 de Octubre, i+12, and Molecular Oncology Unit, CIEMAT, Madrid, Spain. [8]CIBERONC, Madrid, Spain. Guillermo de Velasco, Lucia Trilla-Fuertes, Juan Angel Fresno Vara and Daniel Castellano contributed equally to this work. Correspondence and requests for materials should be addressed to G.V. (email: gdvelasco.gdv@gmail.com)

Urothelial tumors	
Number of patients	58
Age (years)	
≤60	**20(34,5%)**
>60	**38(65,5%)**
Median(IQR)	68(60–71)
Range	45–78
Sex	
Male	51(88%)
Female	7(12%)
pT category	
pT2a	2(3.5%)
pT2b	10(17.3%)
pT3a	27(46.5%)
pT3b	8(13.8%)
pT4a	9(15.5%)
pT4b	1(1.7%)
Missing	1(1.7%)
pN category	
pN0	32(55%)
pN1	14(24%)
pN2	6(10%)
Missing	6(10%)
Highest G grade	
G1-2	8(14%)
G3	44(76%)

Table 1. Study population.

In the last years, proteomics approaches have been incorporated into the study of clinical samples, as a way to complement the information provided by classical factors and genomics. Mass spectrometry-based proteomics have emerged as preferred components of a strategy for discovering diagnostic and prognostic protein biomarkers and as well as new therapeutic targets[8]. These investigations are very encouraging[9,10] and the potential of tumor biomarkers discovery is unclear[11].

Genomics advance in UC has not been translated into molecularly-based biomarker for treatment selection. Since few data is available with proteomics, we aimed to identify whether differentially expressed protein biomarkers in tumor tissue may predict different outcomes.

Results

Study Population. Fifty eight patients with a median age of 68 years (range 45–78 years) were included. Main characteristics are displayed in Table 1. After a median follow up of 38 months, 34 (58.6%) patients relapsed and 35 (60.4%) had died. Median follow-up of all patients was 34 months (range 3–114 months). Median distant disease free survival was 27.7 (27.2–45.1, 95%CI). Five- years-distant relapse free survival was: 75% in stage I/II, 45% in stage III and 25% in stage IV.

Protein preparation and mass spectrometry analysis. After mass spectrometry (MS) workflow, 58 urothelial tumors were analyzed. Raw data normalization was performed as described previously[12]. 4,405 protein groups were identified using Andromeda, of which 1,453 presented at least two unique peptides and detectable expression in at least 75% of the samples. No decoy protein passed through these additional filters.

Protein expression analyses of urothelial tumors and identification of new molecular subtypes. Proteomics data from 58 MIUC tumors were analyzed using sparse k-means and random-forest in order to establish a consistent classification of our samples. Using these approaches, two different molecular groups were identified on the basis of 34 proteins differentially expressed between both groups (Supplementary Figure 1, Supplementary Table 1). From those, 20 proteins have higher expression in group 1, including EHD2, FLNA and TNS1. Gene ontology analyses showed that these proteins are mainly related with focal adhesion and extracellular matrix. On the other hand, 14 proteins have higher expression in group 2, including HSBP1. Gene ontology analyses showed that these proteins are mainly related with transcription processes and immune response. Group 1 showed better prognosis than Group 2, although these differences were not significant (Fig. 1). Contingency analyses showed that these two groups are independent of clinical factors such as stage, tumor size and lymph node status.

Network construction and functional node assignation. Protein expression data from all samples were used in the probabilistic graphical models analyses, with no other *a priori* information. The resulting graph was processed (Fig. 2) looking for a functional structure, i.e., if the proteins included in each branch of the tree

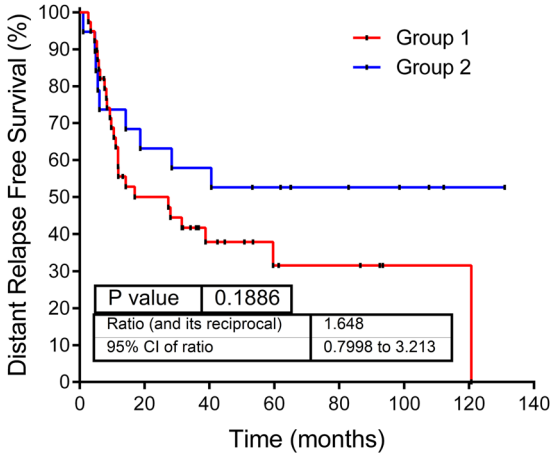


Figure 1. Kaplan–Meier survival curves obtained from high/low risk groups originated in our classification.

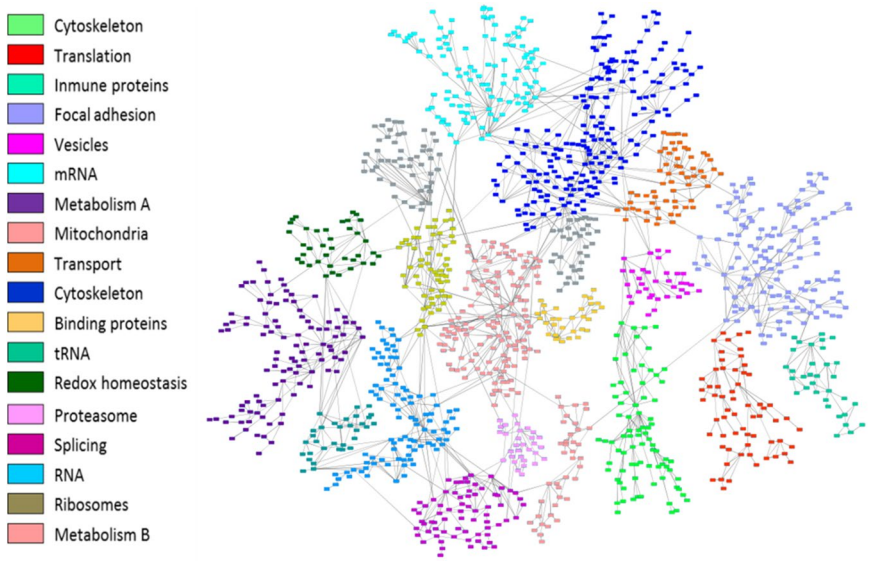


Figure 2. Probabilistic graphical model analysis unravels the functional organization of proteins in MIUC based on correlation. Grey nodes are nodes without any majority function assigned.

had some relationship regarding their function, as previously described[12]. Thus, we divided our graph into eighteen branches, and performed gene ontology analyses. The structure of the probabilistic graphical model had a strong biological function basis. The next step was to calculate the activity for each branch with a specific biological function, i.,e., a functional node, as previously described[12] (Supplementary Figure 2). Once calculated, we evaluated the prognostic value of each functional node activity in MIUC. Focal adhesion functional node activity splits the population into two groups with different prognosis ($p = 0.0241$, $HR = 0.44$ $IC95 = 0.234$ to 0.899) (Fig. 3). Afterwards, we assessed the differences in the functional nodes activities between Group 1 and Group 2 using class comparison analyses. Twelve nodes showed significant different activity between both groups. Focal adhesion, two cytoskeleton nodes, tRNA, ribosomes and metabolism A & B functional nodes showed increased activity in Group 1 tumors, whereas vesicles, transport, proteasome, RNA and splicing nodes showed increased activity in Group 2 tumors (Supplementary Figure 2).

Focal adhesion functional node. Focal adhesion functional node includes twenty six proteins related with extracellular matrix and focal adhesion. COL1A1, SOD3, COL6A1, COL6A2, CAPN2, MSN, STOM, PRELP, NID2, DAG1, LPP and GPI are highly expressed in group 1 while SFN and HDLBP are highly expressed in group 2 ($p < 0.05$). Overall, functional activity of this node is higher in group 1. In addition, this functional node has prognostic value in our cohort.

Development of a prognostic protein signature in MIUC. 66 proteins were found to be associated with recurrence risk in MIUC (Supplementary Table 2). A recurrence signature was developed as previously described[13]. Six proteins of these 66 were included in the prognostic signature: ANXA1 (Annexin A1), BGN (Biglycan), IGFBP7 (Insulin Like Growth Factor Binding Protein 7), ISLR (Immunoglobulin Superfamily

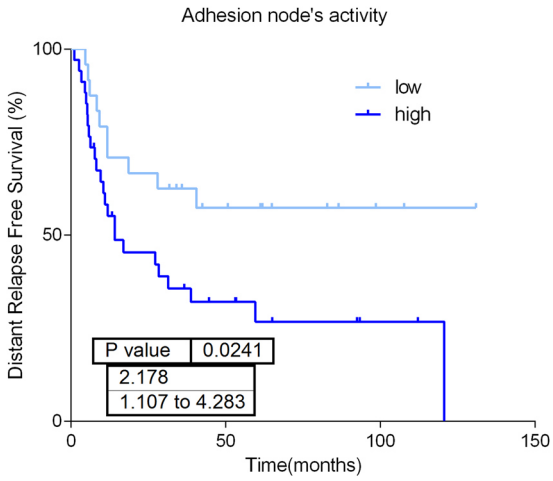


Figure 3. Focal adhesion node's activity has prognostic value (p-value = 0.0241, HR = 2.178, IC95 = 1.107 to 4.283).

Containing Leucine-Rich Repeat), MDP1 (Magnesium-Dependent Phosphatase 1) and PLS3 (Plastin 3). The recurrence protein score split our population into two risk groups with different five year distant relapse free survival: 70% vs. 20% (HR 3.53 95% CI 1.8–6.7; [$p < 0.001$]) (Fig. 4). The association between the score and DRFS was similar for patients with stage III and those with stage IV (Fig. 4). These results were verified using gene expression data from the MD Anderson cohort[4]. In this population, the 6 proteins predictor identifies two populations with different relapse risk (HR 2.10 95% CI 0.2–1.1; [$p = 0.04$]) (Supplementary Figure 3).

Functional proteomics add prognostic information to prognostic signature. Information provided by the prognostic signature is complementary to the prognostic information provided by focal adhesion functional node activity signature and both predictors combined establish four different classes into the population with different relapse risk (p-value = 0.0003) (Fig. 5). Univariate analyses of clinical (stage, tumor size and nodal involvement) and proteomics-based variables (6 protein signature and focal adhesion node activity) showed that focal adhesion node activity (p-value = 0.013), 6 proteins signature (p-value = 0.002) and tumor size (p-value = 0.033) have prognostic value in MIUC. Multivariate analyses showed that both focal adhesion (p-value = 0.011) and 6 proteins signature (p-value = 0.020) has independent prognostic value (Table 2).

Discussion

The principal aim of the study was to establish a molecular classification and survival prediction in MIUC based on proteome analysis since bladder cancer classifications have generally been based on histology features[14] and genomics have yet to be implemented in the clinic. In the clinical practice, there seems to be different groups of patients beyond pathological characteristics. Some patients, even with positive lymph nodes, may never relapse after surgery. However, other subset of patients with apparent favorable features may become metastatic. Therefore it is clear that stratification using the current system is inadequate to satisfactorily differentiate prognosis. Consequently, it is necessary to characterize MIUC patients in accordance to prognostic evolution and molecular features. The proposal of new classifications and characterization of MIUC could lead to further stratification of MIUC tumors and may drive treatment selection.

Our proteomics pipeline allowed us detecting 4,405 proteins in 58 FFPE MIUC samples. We identified groups in our protein data using sparse k-means and confirmed its consistency by random forest, supporting that different molecular subgroups exist in MIUC. Sparse k-means classification is based on 34 proteins, most of them related with focal adhesion. These two molecular groups provide additional information to clinical parameters. As we demonstrated in previous works, probabilistic graphical models using protein expression data allow characterizing differences in biological processes and pathways between groups of patients[12,15]. We were able to establish different functional nodes according to biological functions. The analysis identified 18 different functional nodes, 12 of them monitoring eleven biological processes showed differential activity between the prognostic groups previously established. These results confirm that this approach is valid to study the differential activity of biological functions between tumor groups[12].

Group 1 showed higher expression of some proteins related with focal adhesion and extracellular matrix. Specifically, some of these proteins have been related with epithelial-to-mesenchymal transition (EMT) markers such as EH domain containing 2 (EHD2), which can inhibit metastasis by regulating cadherins[16] or tensin 1 (TNS1), involved in focal adhesion. Low levels of TNS1 have been associated with worsening-free survival in non-muscle invasive bladder cancer[17]. Additionally, filamin A, a downstream effector of mTORC2, plays an important role in motility and invasion[18]. Additionally, we showed an increased activity of biological processes in Group 1, such as Cytoskeleton and Focal Adhesion, Metabolism and tRNA and ribosomes. It is noteworthy that related nodes, such as Cytoskeleton and Focal Adhesion nodes, and also tRNA and ribosomes nodes showed a similar behaviour, showing consistency for obtained biological information. Metabolism A node includes proteins related with negative regulation of protein metabolic process whereas Metabolism B node included proteins

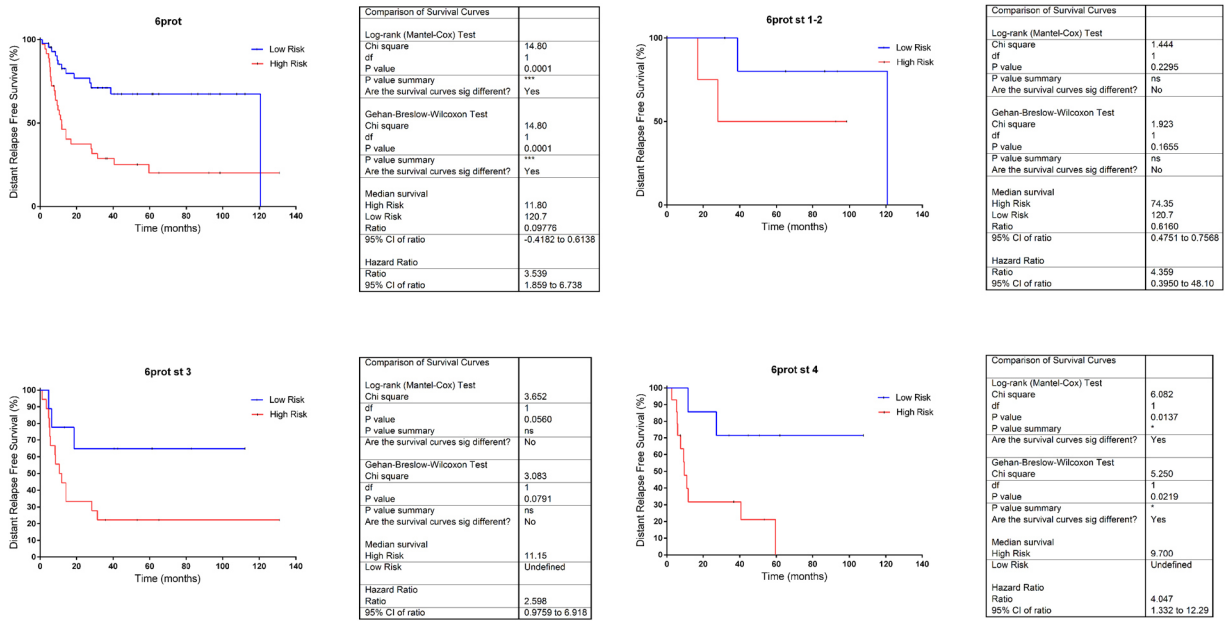


Figure 4. Prognostic signature composed by 6 proteins. A. All data. B. Stage 1–2. C. Stage 3 D. Stage 4.

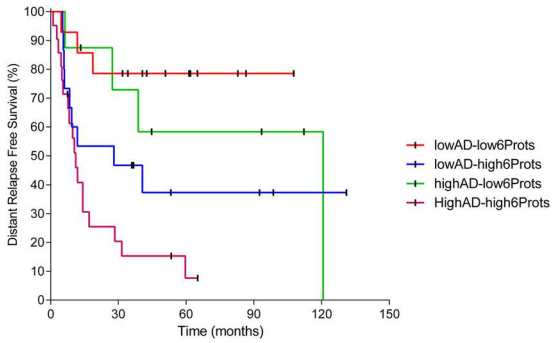


Figure 5. Kaplan–Meier curve curves showing overall survival based on 6 protein signature merged with focal adhesion node activity signature (p-value = 0.0003).

Multivariate analysis				
			95.0% IC para Exp(B)	
	Sig.	Exp(B)	Inferior	Superior
6prots	0.020	3.486	1.217	9.981
Adhesion Node	0.011	3.029	1.287	7.130
Stage	0.840	1.086	0.489	2.412
Size	0.452	0.910	0.711	1.164
N	0.747	1.150	0.492	2.687

Table 2. Multivariate analysis.

related to glycolysis and pyruvate metabolism, involved in generation of precursor metabolites and energy. All together, these results suggest that Group 1 have lower metastatic potential and specific features regarding metabolism and protein synthesis when compared with Group 2.

On the other hand, several proteins showed higher expression in group 2. Some immune proteins such as HSBP1 (heat shock factor binding protein 1) were associated with a decreased immune activity which may have therapeutic implications[19]. Additionally, group 2 showed increased activity in Vesicles, Transport, Proteasome, Splicing and RNA nodes. Again, we found coherence in the biological information, as long as nodes with comparable function showed similar behavior. These results suggest differences regarding intracellular trafficking, RNA processing and Proteasome activities when comparing new defined groups.

The differences in biological functions, after proper validation, could lead to develop specific treatments in concrete groups of patients. For instance, differences in metabolism could be targeted with 2-D-deoxy-glucose or metformin[20], which are being currently tested in clinical trials for breast cancer treatment. On the other hand, proteasome targeting drugs have demonstrated therapeutic value in multiple myeloma treatment[21].

In this study we show that the discovery of proteins as prognostic biomarkers is feasible using FFPE samples and proteomics. Indeed, we were able to identify a six protein-signature with prognosis value independently of stage, size and lymph node status. Proteins contained in this predictor are involved in multiple processes. ANXA1, a membrane-located protein, has been related with prognosis in breast cancer[22]. BGN is a protein involved in inflammation processes[23]. Niedworok *et al*.[24] suggested that biglycan is an endogenous inhibitor of bladder cancer cell proliferation and its high expression is associated with good prognosis. PLS3 was proposed as biomarker for breast cancer prognosis[25]. To our knowledge, no previous information about MDP1, IGFBP7 and ISLR role in cancer disease has been previously reported. The prognostic value of the 6-protein signature was validated using gene expression data from another cohort.

Probabilistic graphical models allow to compare biological functions between groups but also, to build prognostic signatures. Focal adhesion functional node activity had prognostic value and split population in low and high risk of relapse. Strikingly, prognostic information provided by a traditional protein signature was complementary to information provided by focal adhesion functional node activity signature, and also to the prognostic information provided by clinical factors, as shown in the multivariate analysis. Merging these molecular features, it is feasible to establish four different risk populations. These results confirm that functional approaches could provide additional information to traditional gene/protein-centered analyses.

Our study has some limitations. Technically, proteomics provide less information when compared with genomics, thus an improvement in number of detected proteins is still necessary. On the other hand, peptide and protein identification relies in statistical parameters. Due to this, we applied strict filters for peptides/proteins selection, in order to avoid false detections, ensuring that proteins with the highest confidence in both identification and quantification are selected for analyses. Finally, although a meta-validation has been performed, these results should be validated in additional cohorts to evaluate the 6-protein signature robustness and the functional differences between new defined molecular groups. Other limitations of this study include the relatively small sample size and there may be other bias that could affect outcomes. We believe that our findings serve as important hypothesis generating findings that can be explored in future studies.

In conclusion, our approach, combining proteomics and probabilistic graphical models allow the integration of different levels of molecular information that can improve MIUC molecular characterization. We were able to differentiate two different molecular groups from our proteomics data, with different functional features that may represent new therapeutic opportunities for bladder cancer treatment. Moreover, we defined a 6 protein-signature that can predict the outcome of MIUC patients and we identified a functional node with prognosis value in MIUC, adding prognostic information to the prognostic 6-protein signature and to clinical factors.

Methods

Patient's characteristics and samples selection. Patients treated at University Hospital 12 de Octubre (Madrid, Spain) were included if they had histologically documented (TNM staging[26], T1-T4a and any N, M0) urothelial carcinoma (including of the renal pelvis, ureter, urinary bladder, or urethra). In total, 58 patients who underwent curative surgical resection between 2006 and 2012 were selected. FFPE samples were retrieved from I + 12 Biobank (RD09/0076/00118). Samples were reviewed by a genitourinary pathologist and included if cases had at least 50% of urothelial tumor cells and were invasive in the muscularis propria. The study was approved by independent review board and Ethical Committee of Hospital Universitario 12 de Octubre. All experiments were performed in accordance with relevant guidelines and regulations. Informed consent was obtained from all participants before starting treatment.

Liquid chromatography - mass spectrometry shotgun analysis. Proteins were extracted from FFPE samples as previously described[27]. Mass spectrometry analysis was performed on a QExactive mass spectrometer coupled to a nano EasyLC 1000 (Thermo Fisher Scientific). Solvent composition at the two channels was 0.1% formic acid for channel A and 0.1% formic acid, 99.9% acetonitrile for channel B. For each sample 2 μL of peptides were loaded on a self-made column (75 μm × 150 mm) packed with reverse-phase C18 material (ReproSil-Pur 120 C18-AQ, 1.9 μm, Dr. Maisch GmbH) and eluted at a flow rate of 300 nL/min by a gradient from 2 to 35% B in 80 min, 47% B in 4 min and 98% B in 4 min. Samples were acquired in a randomized order. The mass spectrometer was operated in data-dependent mode (DDA), acquiring a full-scan MS spectra (300–1700 m/z) at a resolution of 70000 at 200 m/z after accumulation to a target value of 3000000, followed by HCD (higher-energy collision dissociation) fragmentation on the twelve most intense signals per cycle. HCD spectra were acquired at a resolution of 35000 using normalized collision energy of 25 and a maximum injection time of 120 ms. The automatic gain control (AGC) was set to 50000 ions. Charge state screening was enabled and singly and unassigned charge states were rejected. Only precursors with intensity above 8300 were selected for MS/MS (2% underfill ratio). Precursor masses previously selected for MS/MS measurement were excluded from further selection for 30 s, and the exclusion window was set at 10 ppm. The samples were acquired using internal lock mass calibration on m/z 371.1010 and 445.1200.

Protein identification and label free protein quantification. The acquired raw MS data were processed by MaxQuant (version 1.5.2.8), followed by protein identification using the integrated Andromeda search engine. Spectra were searched against a forward Swiss Prot-human database, concatenated to a reversed decoyed fasta database and common protein contaminants (NCBI taxonomy ID9606, release date 2014-05-06). Carbamidomethylation of cysteine was set as fixed modification, while methionine oxidation and N-terminal

protein acetylation were set as variable. Enzyme specificity was set to trypsin/P allowing a minimal peptide length of 7 amino acids and a maximum of two missed-cleavages. Precursor and fragment tolerance was set to 10 ppm and 20 ppm, respectively for the initial search. The maximum false discovery rate (FDR) was set to 0.01 for peptides and 0.05 for proteins. Label free quantification was enabled and a 2 minutes window for match between runs was applied. The re-quantify option was selected. For protein abundance the intensity was used, corresponding to the sum of the precursor intensities of all identified peptides for the respective protein group.

Sparse k-means classification. Sparse k-means was used to establish differential groups between samples. Classification consistency was tested using random forest. An analysis with the Consensus Clustering algorithm[28], applied on the data containing the variables that were selected by the sparse K-means method[29], has provided an optimum classification into two subtypes in previous studies[30].

Functional network construction. Network construction was performed using probabilistic graphical models compatible with high dimensional data using correlation as associative method as previously described[12]. In order to identify functional nodes in the networks we split them in several branches and we used Gene Ontology analysis to assign a majority function to each node. Activity measurement was calculated by the mean expression of all the proteins of each node related with the assigned node function.

Gene-Ontology Analysis. Protein to Gene Symbol conversion was performed using Uniprot and DAVID[31]. Gene Ontology Analysis was also done in DAVID selecting "*Homo sapiens*" background and GOTERM-FAT, Biocarta, KEGG and Panther databases.

Protein signature construction. We computed a statistical significance level for each protein based on a univariate proportional hazards model with the aim of identifying proteins whose expression were significantly related to the distant metastasis-free survival (DMFS) as described previously[13]. Leave-one-out cross-validation was used to evaluate the predictive accuracy of the profiles. The cutoff point was established *a priori* and to test the statistical significance, the p-value of the log-rank test statistic for the risk groups was evaluated using 1000 random permutations. Analyses were performed in BRB-ArrayTools v4_2_1 and R v3.2.4[32]. BRB-ArrayTools has been developed by Dr. Richard Simon and BRB-ArrayTools Development Team.

Prognostic signature meta-validation. With the aim to verify the utility of 6 protein signature, gene expression data from a MD Anderson cohort was used[4]. All probes in dataset for each gene were retrieved. Probes with higher CV were selected when multiple probes were found for a single gene, then expression values of each gene were z-score transformed as previously described[15]. To apply protein expression based signatures to gene expression values, per-gene normalization was applied as previously described[13].

Statistical analyses. Statistical analyses (class comparisons contingency analyses, etc.), were performed using GraphPad Prism v6 and Cytoscape[33]. Univariate and multivariate Cox regression models were performed using IBM SPSS Statistics. All p-values where two-sided, and $p < 0.05$ was considered statistically significant.

References

1. Ferlay, J. *et al.* Cancer incidence and mortality worldwide: sources, methods and major patterns in GLOBOCAN 2012. *Int J Cancer* **136**, E359–386, https://doi.org/10.1002/ijc.29210 (2015).
2. Collaboration, A. B. C. A. M.-a. Neoadjuvant chemotherapy in invasive bladder cancer: update of a systematic review and meta-analysis of individual patient data advanced bladder cancer (ABC) meta-analysis collaboration. *Eur Urol* **48**, 202–205; discussion 205–206, https://doi.org/10.1016/j.eururo.2005.04.006 (2005).
3. Penson, D. F. Re: Effectiveness of Adjuvant Chemotherapy for Locally Advanced Bladder Cancer. *J Urol* **196**, 352–354, https://doi.org/10.1016/j.juro.2016.05.032 (2016).
4. Choi, W. *et al.* Identification of distinct basal and luminal subtypes of muscle-invasive bladder cancer with different sensitivities to frontline chemotherapy. *Cancer Cell* **25**, 152–165, https://doi.org/10.1016/j.ccr.2014.01.009 (2014).
5. Kim, J. *et al.* Invasive Bladder Cancer: Genomic Insights and Therapeutic Promise. *Clin Cancer Res* **21**, 4514–4524, https://doi.org/10.1158/1078-0432.CCR-14-1215 (2015).
6. Stadler, W. *et al.* Randomized trial of p53 targeted adjuvant therapy for patients with organ-confined node-negative urothelial bladder cancer. *Journal of Clinical Oncology* **27**, 5017–5017 (2009).
7. George, B. *et al.* p53 gene and protein status: the role of p53 alterations in predicting outcome in patients with bladder cancer. *J Clin Oncol* **25**, 5352–5358, https://doi.org/10.1200/JCO.2006.10.4125 (2007).
8. Hanash, S. Disease proteomics. *Nature* **422**, 226–232, https://doi.org/10.1038/nature01514 (2003).
9. Marko-Varga, G. *et al.* Personalized medicine and proteomics: lessons from non-small cell lung cancer. *J Proteome Res* **6**, 2925–2935, https://doi.org/10.1021/pr070046s (2007).
10. Pastwa, E., Somiari, S. B., Czyz, M. & Somiari, R. I. Proteomics in human cancer research. *Proteomics Clin Appl* **1**, 4–17, https://doi.org/10.1002/prca.200600369 (2007).
11. Rifai, N., Gillette, M. A. & Carr, S. A. Protein biomarker discovery and validation: the long and uncertain path to clinical utility. *Nat Biotechnol* **24**, 971–983, https://doi.org/10.1038/nbt1235 (2006).
12. Gámez-Pozo, A. *et al.* Vol. 75 2243–2253 (Cancer Res, 2015).
13. Sánchez-Navarro, I. *et al.* An 8-gene qRT-PCR-based gene expression score that has prognostic value in early breast cancer. *BMC Cancer* **10**, 336, https://doi.org/10.1186/1471-2407-10-336 (2010).
14. Humphrey, P. A., Moch, H., Cubilla, A. L., Ulbright, T. M. & Reuter, V. E. The 2016 WHO Classification of Tumours of the Urinary System and Male Genital Organs-Part B: Prostate and Bladder Tumours. *Eur Urol* **70**, 106–119, https://doi.org/10.1016/j.eururo.2016.02.028 (2016).
15. Gámez-Pozo, A. *et al.* Functional proteomics outlines the complexity of breast cancer molecular subtypes. *Scientific Reports* **7**, 10100, https://doi.org/10.1038/s41598-017-10493-w (2017).
16. Shi, Y. *et al.* Decreased expression and prognostic role of EHD2 in human breast carcinoma: correlation with E-cadherin. *J Mol Histol* **46**, 221–231, https://doi.org/10.1007/s10735-015-9614-7 (2015).
17. Fahmy, M. *et al.* Relevance of the mammalian target of rapamycin pathway in the prognosis of patients with high-risk non-muscle invasive bladder cancer. *Hum Pathol* **44**, 1766–1772, https://doi.org/10.1016/j.humpath.2012.11.026 (2013).

18. Chantaravisoot, N., Wongkongkathep, P., Loo, J. A., Mischel, P. S. & Tamanoi, F. Significance of filamin A in mTORC2 function in glioblastoma. *Mol Cancer* **14**, 127, https://doi.org/10.1186/s12943-015-0396-z (2015).
19. Yashin, D. V. *et al*. The heat shock-binding protein (HspBP1) protects cells against the cytotoxic action of the Tag7-Hsp70 complex. *J Biol Chem* **286**, 10258–10264, https://doi.org/10.1074/jbc.M110.163436 (2011).
20. *ClinicalTrials.gov A service of the U.S National Institutes of Health*.
21. Palumbo, A. *et al*. Daratumumab, Bortezomib, and Dexamethasone for Multiple Myeloma. *N Engl J Med* **375**, 754–766, https://doi.org/10.1056/NEJMoa1606038 (2016).
22. Bhardwaj, A. *et al*. Annexin A1 Preferentially Predicts Poor Prognosis of Basal-Like Breast Cancer Patients by Activating mTOR-S6 Signaling. *PLoS One* **10**, e0127678, https://doi.org/10.1371/journal.pone.0127678 (2015).
23. Hsieh, L. T., Nastase, M. V., Zeng-Brouwers, J., Iozzo, R. V. & Schaefer, L. Soluble biglycan as a biomarker of inflammatory renal diseases. *Int J Biochem Cell Biol* **54**, 223–235, https://doi.org/10.1016/j.biocel.2014.07.020 (2014).
24. Niedworok, C. *et al*. Inhibitory role of the small leucine-rich proteoglycan biglycan in bladder cancer. *PLoS One* **8**, e80084, https://doi.org/10.1371/journal.pone.0080084 (2013).
25. Ueo, H. *et al*. Circulating tumour cell-derived plastin3 is a novel marker for predicting long-term prognosis in patients with breast cancer. *Br J Cancer* **112**, 1519–1526, https://doi.org/10.1038/bjc.2015.132 (2015).
26. Edge, S. *et al*. *Urinary bladder*. 7th edition, New York, Springer edn, 497-505 (AJCC Cancer Staging Manual, 2010).
27. Gámez-Pozo, A. *et al*. Shotgun proteomics of archival triple-negative breast cancer samples. *Proteomics Clin Appl* **7**, 283–291, https://doi.org/10.1002/prca.201200048 (2013).
28. Monti, S., Tamayo, P., Mesirov, J. & Golub, T. Consensus Clustering: A Resampling-Based Method for Class Discovery and Visualization of Gene Expression Microarray Data. *Machine learning* **52**, 91–118 (2003).
29. Witten, D. M. & Tibshirani, R. A framework for feature selection in clustering. *J Am Stat Assoc* **105**, 713–726, https://doi.org/10.1198/jasa.2010.tm09415 (2010).
30. Rousseeuw, P. J. Silhouettes. A graphical aid to the interpretation and validation of cluster analysis. Journal of Computational and Applied Mathematics. *Journal of Computational and Applied Mathematics* **20**, 53–65 (1987).
31. Huang, d. W., Sherman, B. T. & Lempicki, R. A. Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nat Protoc* **4**, 44–57, https://doi.org/10.1038/nprot.2008.211 (2009).
32. (Vienna, Austria. R Foundation for Stattistical Computing, 2013).
33. Shannon, P. *et al*. Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Res* **13**, 2498–2504, https://doi.org/10.1101/gr.1239303 (2003).

Acknowledgements

I + 12 Biobank (RD09/0076/00118) is integrated in the Spanish Hospital Biobank Network (RetBioH; www.redbiobancos.es). LT-F is supported by Spanish Economy and Competitiveness Ministry (DI-15-07614). This work was partially supported by a grant from the FMM (2012-0085) to GdV and by FEDER cofounded MINECO (SAF2015-66015-R) and ISCIII grants (RD12/0036/0009, PIE15/00076, and CB/16/00228 to JMP.

Author Contributions

Conception and design, obtaining funding: A.G.P., J.A.F.V., D.C., G.D.V.; acquisition of data: G.R.A., I.O., F.V., J.M.S., D.C., M.U., G.D.V.; drafting of the manuscript: L.T.F., A.G.P., G.D.V.; Experimental procedures: R.L.V., A.G.P., L.T.F., A.Z.M.; revision of the manuscript: all authors; statistical analysis: L.T.F., A.G.P.; G.P.V., J.M.A., H.N., G.P.V.; supervision: A.G.P., J.A.F.V., D.C., H.N.

Additional Information

Supplementary information accompanies this paper at https://doi.org/10.1038/s41598-017-15920-6.

Competing Interests: J.A.F.V. and A.G.-P. are shareholders in Biomedica Molecular Medicine S.L. L.T.-F. is an employee of Biomedica Molecular Medicine S.L. The other authors declare that they have no competing interests.

Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

www.oncotarget.com Oncotarget, 2018, Vol. 9, (No. 45), pp: 27586-27594

Research Paper: Immunology

Probabilistic graphical models relate immune status with response to neoadjuvant chemotherapy in breast cancer

Andrea Zapater-Moros[1], Angelo Gámez-Pozo[1,2], Guillermo Prado-Vázquez[1], Lucía Trilla-Fuertes[2], Jorge M. Arevalillo[3], Mariana Díaz-Almirón[4], Hilario Navarro[3], Paloma Maín[5], Jaime Feliú[6,7], Pilar Zamora[6], Enrique Espinosa[6,7] and Juan Ángel Fresno Vara[1,2,7]

[1] Molecular Oncology & Pathology Laboratory, Institute of Medical and Molecular Genetics-INGEMM, La Paz University Hospital-IdiPAZ, Madrid, Spain

[2] Biomedica Molecular Medicine SL, Madrid, Spain

[3] Operational Research and Numerical Analysis, National Distance Education University, Madrid, Spain

[4] Biostatistics Unit, La Paz University Hospital-IdiPAZ, Madrid, Spain

[5] Department of Statistics and Operations Research, Faculty of Mathematics, Complutense University of Madrid, Madrid, Spain

[6] Medical Oncology Service, La Paz University Hospital-IdiPAZ, Madrid, Spain

[7] CIBERONC, Madrid, Spain

***Correspondence to**: Juan Ángel Fresno Vara, **email**: juanangel.fresno@salud.madrid.org*

Keywords: breast cancer; neoadjuvant chemotherapy; molecular subtypes; probabilistic graphical models; immune status; Immunology

***Received**: December 21, 2017* ***Accepted**: May 08, 2018* ***Published**: June 12, 2018*

ABSTRACT

Breast cancer is the most frequent tumor in women and its incidence is increasing. Neoadjuvant chemotherapy has become standard of care as a complement to surgery in locally advanced or poor-prognosis early stage disease. The achievement of a complete response to neoadjuvant chemotherapy correlates with prognosis but it is not possible to predict who will obtain an excellent response. The molecular analysis of the tumor offers a unique opportunity to unveil predictive factors. In this work, gene expression profiling in 279 tumor samples from patients receiving neoadjuvant chemotherapy was performed and probabilistic graphical models were used. This approach enables addressing biological and clinical questions from a Systems Biology perspective, allowing to deal with large gene expression data and their interactions. Tumors presenting complete response to neoadjuvant chemotherapy had a higher activity of immune related functions compared to resistant tumors. Similarly, samples from complete responders presented higher expression of lymphocyte cell lineage markers, immune-activating and immune-suppressive markers, which may correlate with tumor infiltration by lymphocytes (TILs). These results suggest that the patient's immune system plays a key role in tumor response to neoadjuvant treatment. However, future studies with larger cohorts are necessary to validate these hypotheses.

INTRODUCTION

Breast cancer is the most common neoplasm and the fifth cause of cancer-associated death among women [1]. Estrogen receptor (ER), progesterone receptor (PR), and human epidermal growth factor receptor 2 (HER2) provide a system of classification and clinical diagnosis. Seventy percent of the tumors are hormonal receptor positive, and HER2 overexpression is observed in 15% of cases. ER+ and PR+ tumors respond to endocrine therapy, whereas tumors overexpressing HER2 respond to targeted therapies such as trastuzumab [2, 3]. Tumors negative for

ER, PR and HER2 are known as Triple Negative Breast Cancer (TNBC) and do not respond to the aforementioned therapies.

A molecular classification of breast cancer defined four intrinsic subtypes [4]. Luminal A disease, which accounts for 67% of all tumors, shows high expression of genes related to hormone receptors and low expression of genes related to cell proliferation. Luminal B, HER2-enriched and Basal-like subtypes have a more aggressive phenotype [5] [6, 7].

Neoadjuvant chemotherapy has been increasingly administered to reduce the size of primary tumor, thus increasing the likelihood of breast conservation and enhancing survival [8]. Currently, there is no clinically useful molecular predictor of response to cytotoxic drugs in the neoadjuvant setting. Clinical parameters or the expression of single molecular markers (ie, Bcl-2, p53, MDR-1, and so on) show weak association with response and are not regimen-specific. Molecular subtyping may offer some help, as Luminal B and Basal-like tumors respond better than Luminal A tumors [9], but this is not accurate enough to make clinical decisions. As a consequence, many patients suffer the toxicity of useless neoadjuvant chemotherapy.

This study has been carried out using probabilistic graphical models, providing insights into the molecular biology of tumor response, allowing its use as a predictive model for response. These statistically inferred networks provide a deeper level of biological understanding in two main directions: giving support to previously identified biological observations, and giving new insights regarding novel biological interactions. Moreover, the transcriptional network approach has proven to be useful to unveil transcriptional regulation in breast cancer [10, 11]. The objective of this study was to evaluate differences in gene expression patterns of breast cancer tumors from patients who had undergone neoadjuvant chemotherapy through a Systems Biology perspective.

RESULTS

Patient's characteristics

279 patients with histologically-confirmed primary non-metastatic breast adenocarcinoma from phase II trial (NCT00455533) [12] were included. They all had untreated tumors of at least 2 cm in size (T2-3, N0-3) regardless of hormone receptor or HER2 expression status. Clinical data were obtained from phase II trial (NCT00455533). Patient's clinical characteristics are provided in Table 1. On the basis of ER, PR and HER2 status, 111 tumors patients (39.78%) were ER+ or PR+ and Her2- (ER+ for now on), 28 (10.04%) were HER2+ and 140 (50.18%) were classified as TNBC. Patients received sequential neoadjuvant therapy starting with 4 cycles of doxorubicin/cyclophosphamide (AC), followed by 1:1 randomization to either ixabepilone or paclitaxel. All patients underwent definitive breast surgery 4 to 6

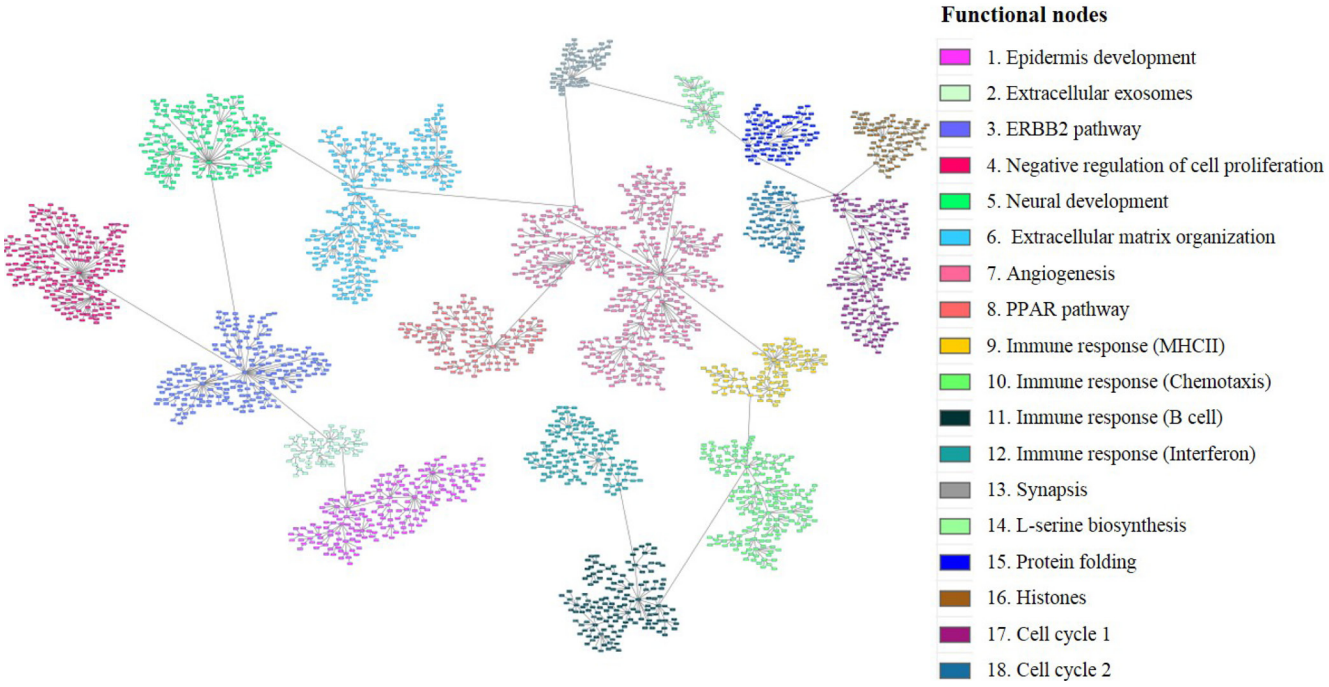


Figure 1: Breast cancer network. Probabilistic graphical model from 279 tumors gene expression data divided in eighteen functional nodes harboring one or two predominant biological functions. Each node (box) represents one gene and each grey line (edges) connects genes with correlated expression.

Table 1: Patient's clinical characteristics

Characteristic	Patients (*n*)	Patients (%)	Characteristics	Patients (*n*)	Patients (%)
Age			**Pathological response**		
Mean age	48.63		CR	40	14.34%
≤50	166	59.50%	PR	161	57.71%
>50	113	40.50%	PD	5	1.79%
Tumor size (T)			SD	64	22.94%
< 2 cm	3	1.08%	Unassigned	9	3.23%
2 - 5 cm	174	62.37%	**ER status**		
> 5 cm	99	35.48%	ER+	108	38.71%
Unassigned	3	1.08%	ER-	171	61.29%
Nodal classification (N)			**PR status**		
N0	122	41.40%	PR+	99	35.48%
N1	136	46.10%	PR-	179	64.16%
N2	30	10.20%	Unknown	1	0.36%
N3	7	2.40%	**HER2 status**		
Neoadjuvant treatment			HER2+	28	10.04%
Ixabepilone	138	49.46%	HER2-	251	89.96%

weeks after the last dose of ixabepilone or paclitaxel, consisting of either a lumpectomy with axillary dissection or modified radical mastectomy. Regarding pathological response, 40 (14.34%) patients achieved a complete response (CR), 161 (57.71%) achieved a partial response (PR), 64 (22.94%) had stable disease (SD) and 5 (1.79%) had progressive disease (PD).

Molecular stratification of tumors

Molecular subtypes were defined by PAM50 assignment [13]. Of the initial 279 patients, 116 (41.58%) patients were classified as Basal-like subtype, 15 patients (5.38%) as HER2+, 66 (23.66%) as Luminal A, 62 (22.22%) as Luminal B, and 15 (5.38%) as Normal-like. Five patients could not be assigned due to Spearmn's rank correlation were not statistically significant for neither of the molecular subgroup centroids. A sub-classification of TNBCs was performed based on Lehmann's classification as previously described [14]; 25 (8.96%) TNBC tumors were Basal-like 1, 83 (29.75%) Basal-like 2 subgroup, 6 (2.15%) Luminal Androgen Receptor, and 26 (9.32%) Mesenchymal.

Breast cancer systems biology

Gene expression data from all tumor samples were used to build a probabilistic graphical model, with no other *a priori* information. The resulting graph was divided in eighteen branches (functional nodes) and a main function was assigned to each node by gene ontology analysis. The structure of the probabilistic graphical model clearly reflected different biological functions (Figure 1) Functional node activities were then calculated as previously showed [10, 15].

Functional structure of response to neoadjuvant chemotherapy

Patients were classified according to pathological response regardless of their tumor molecular subtype to study the response to neoadjuvant chemotherapy. Significant differences between functional node activities were observed in "Immune response (MHCII)" (node 9), "Immune response (B cell)" (node 11) and "Immune response (Interferon)" (node 12) nodes, in which, tumors attaining a complete response had higher activation (Figure 2). Blown up pictures of the genes in the red boxes are provided in Supplementary Figures 1-5. A progressive decrease in the activity of immune functional nodes was seen depending on the response, being higher in tumors obtaining a CR and absent in those having a progression. Additionally, the relationship of immune nodes activities with the pathological response was evaluated using an ordinal logistic regression analysis. This analysis revealed that an increment of one unit in node 9, 11 and 12 activities increased the probability of a favorable response 1.739, 1.435 and 1.629 times respectively. By contrast, one unit increase in the activity of node 10 increased 0.519 times the probability of having an unfavorable response. On the other hand, PD tumors showed higher functional activity in "Cell cycle 1" (node 17) and "Cell cycle 2" (node18), followed by CR tumors.

Functional characterization of molecular subtypes

Patients in the network were further classified according to their molecular subtype (Basal-like, Luminal

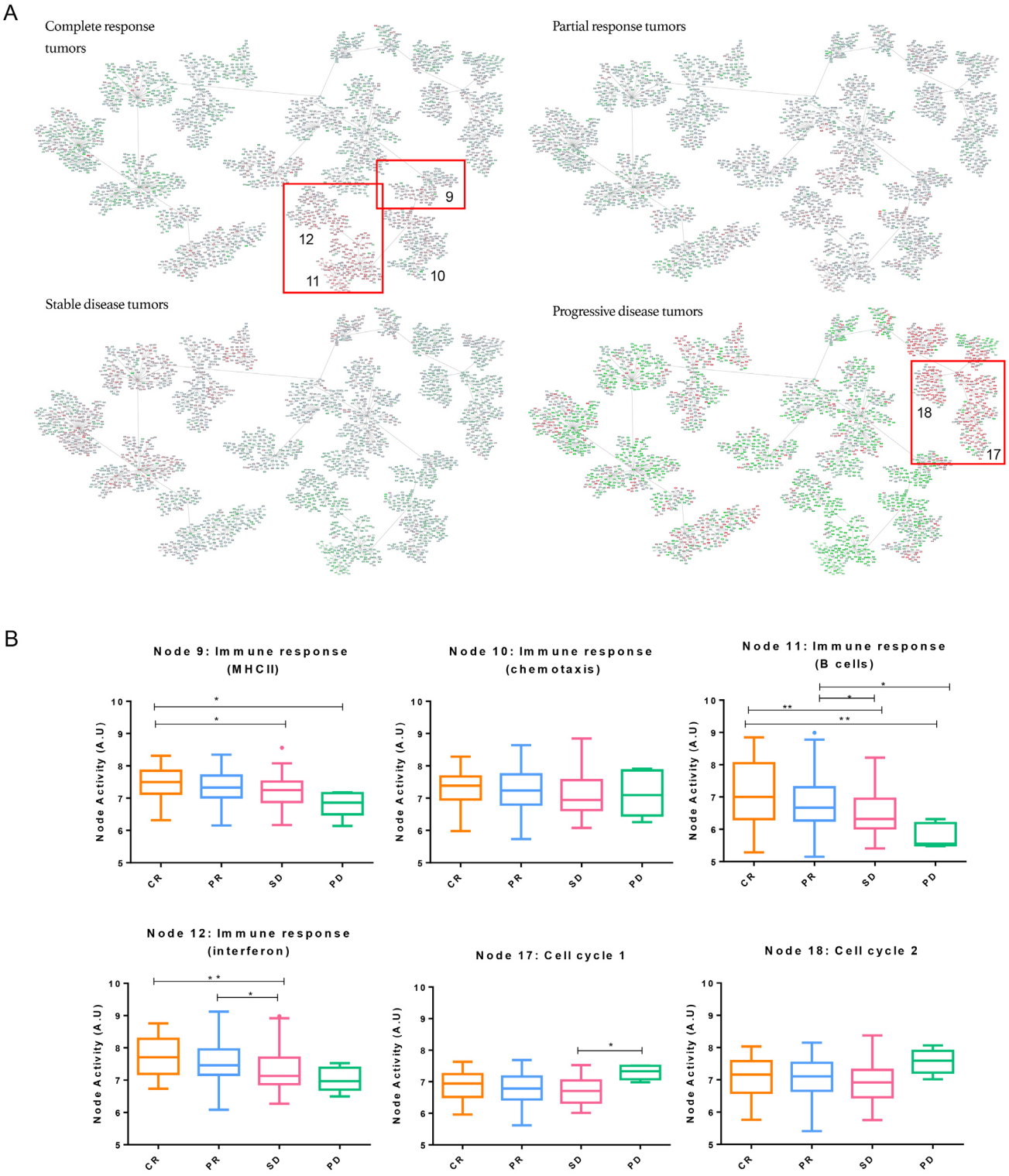


Figure 2: Breast cancer network by pathological response groups. A. Detail of nodes with the highest activity in each of the subgroups. Genes with an expression below 0 were represented in green; genes with an expression around 0 were represented in grey and genes with an expression above zero were represented in red. **B.** Functional node activities differences between pathological response groups: Box-and-whisker plots are Tukey boxplots. All *p*-values were two-sided and $p < 0.05$ was considered statistically significant. *P*-value < 0.05 (*); *p*-value < 0.01(**). A.U: arbitrary units.

A, Luminal B, HER2+ and Normal-like). Basal-like tumors were also classified according to Lehmann's subtypes. "Immune response (MHCII)" (node 9) node activity was higher in Luminal A and Normal-like subtypes while Basal-like tumors showed higher functional node activity in "Immune response (chemotaxis)" (node 10), "Immune response (B cell)" (node 11) and "Immune response (Interferon)" (node 12), as well as in "Cell cycle 1 and 2" (nodes 17 and 18) nodes (Figure 3).

Concerning TNBCs sub-classification, BL2 subtype showed a higher functional activity in "Immune response" (nodes 9, 10, 11 and 12) nodes whereas it was observed higher "Cell cycle 1 and 2" (nodes 17 and 18) nodes activities in BL1 tumors.

In order to evaluate the functional implications between molecular subtypes and response to neoadjuvant therapy, data from patients of the same molecular subtype were mean centred and analysed independently. Luminal A group included no PD, whereas only one PD was found in Luminal B group, and was excluded from this analysis. Normal-like and HER2+ tumors were insufficient to perform subsequent analyses.

Concerning Luminal A and Luminal B subtypes, "Immune response (MHCII)" (node 9), "Immune response (chemotaxis)" (node 10), "Immune response (B cell)" (node 11) and "Immune response (Interferon)"

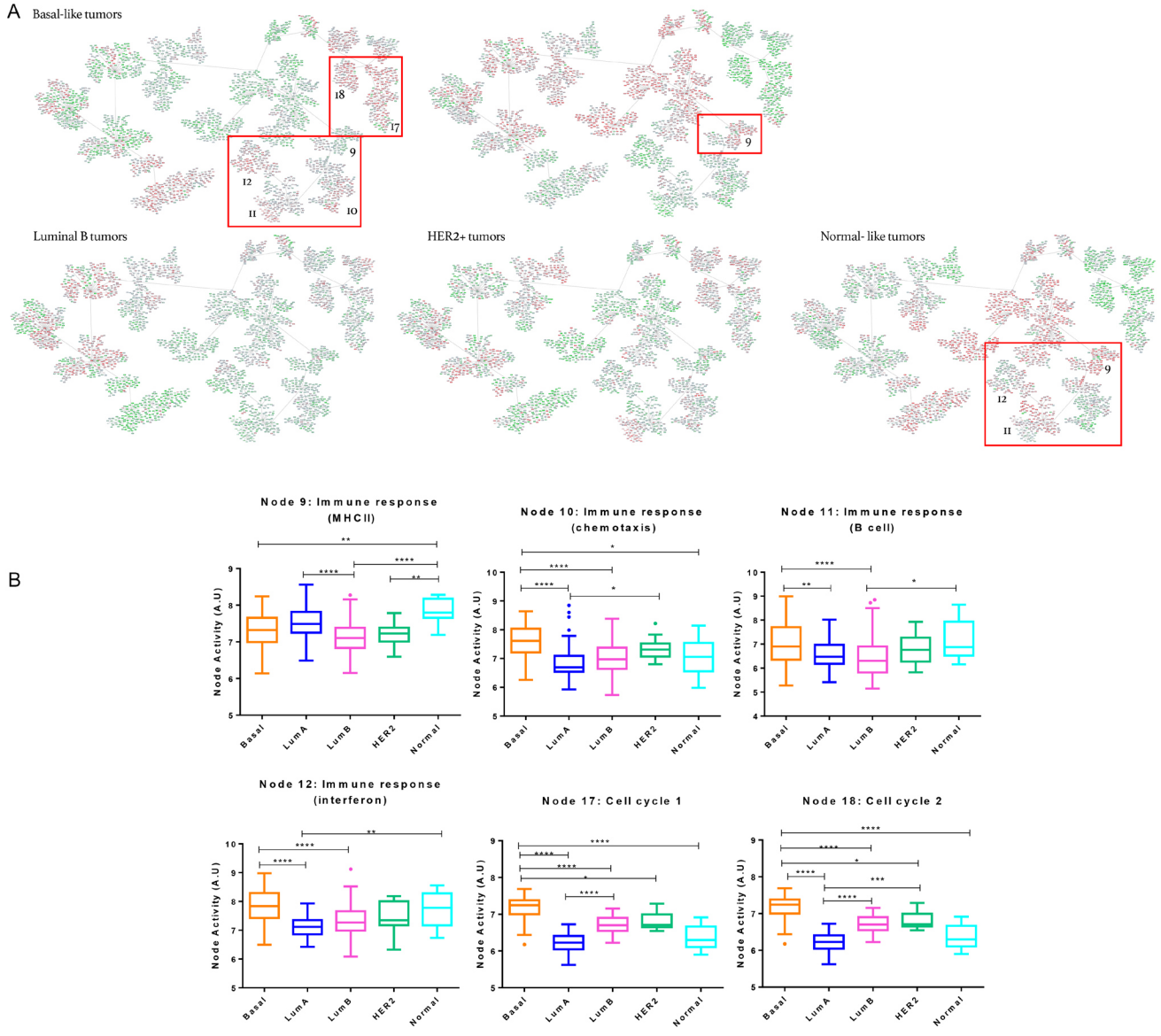


Figure 3: Breast cancer network by breast cancer molecular subtypes. A. Detail of nodes with the highest activity in each of the subgroups. Genes with an expression below 0 were represented in green; genes with an expression around 0 were represented in grey and genes with an expression above zero were represented in red. **B.** Functional node activities differences between molecular subtypes: Box-and-whisker plots are Tukey boxplots. All *p*-values were two-sided and $p < 0.05$ was considered statistically significant. *P*-value < 0.05 (*); *p*-value < 0.01(**); *P*-value< 0.001 (***); *P*-value <0.0001 (****). A.U: arbitrary units.

(node 12) functional nodes activities were higher in tumors attaining a CR, although differences were not statistically significant. As in the case of Basal-like tumors, a progressive decrease of activity in these nodes was observed from CR to SD.

In Basal-like subtype, tumors attaining a CR showed significant differences in "Immune response (B cell)" (node 11) and "Immune response (Interferon)" (node 12) nodes activities while "Immune response (MHCII)" (node 9) node activity was higher in tumors showing a PR. PD tumors showed a higher functional node activity in Cell cycle 1" (node 17) and Cell cycle 2" (node 18). Regarding TNBC, in the BL1 subtype, the activity of nodes "Immune response (B cell)" (node 11) and "Immune response (Interferon)" (node 12) was higher in CR than in PR. "Immune response (MHCII)" (9), "Immune response (chemotaxis)" (node 10) node activities also were higher in CR tumors but without statistics differences. In the BL2 subtype, CR tumors showed a significant higher activity in "Immune response (MHCII)" (node 9) and "Immune response (B cell)" (node 11). However, SD tumors showed a higher activity in "Immune response (chemotaxis)" (node 10). In Mesenchymal subtype, CR tumors showed higher activity in "Immune response (MHCII)" (node 9) and "Immune response (B cell)" (node 11) nodes but without being statistically significant.

Three separate probabilistic graphical models were built for Basal-like, Luminal-A and Luminal-B subtypes. As in the global network, tumors experiencing a CR had an increased activity of immune response-related nodes, although differences were not statistically significant.

Immunological markers

Markers of tumor-infiltrating lymphocytes were analysed to further characterize response according to the immune status. For that, patients were separated according to their pathological response and marker gene expression levels were compared between groups. Tumors obtaining a CR showed significantly higher expression levels of cell lineage markers (CD4, CD8 and CD20) as well as immune-activating (IGKC, CXCL9, CCL5, CXCL13) and immunosuppressive markers (IDO1, PD1) compared to the rest of tumors (Figure 4).

DISCUSSION

In this work, a gene expression-based probabilistic graphical model analysis of breast cancer showed that

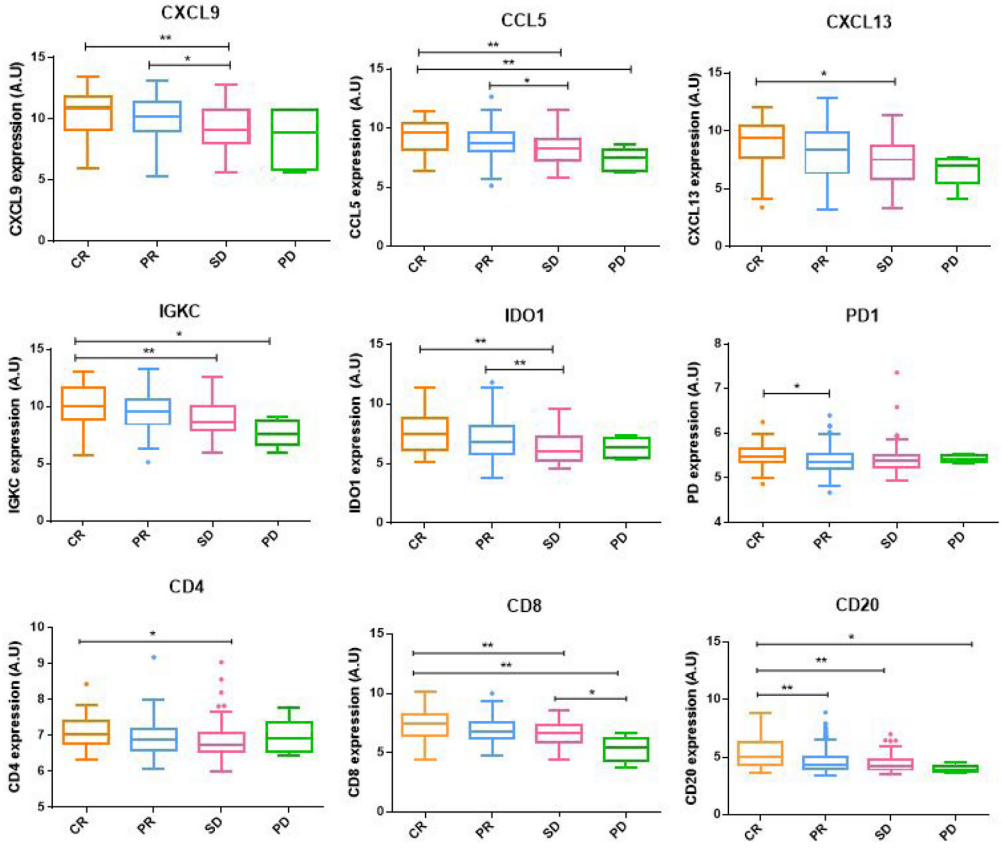


Figure 4: Immunological markers expression. Immune-activating, immunosuppressive and cell lineage markers gene expression between pathological response groups. Box-and-whisker plots are Tukey boxplots. All *p*-values were two-sided and $p < 0.05$ was considered statistically significant. *P*-value < 0.05 (*); *p*-value < 0.01(**); *p*-value < 0.001 (***); *p*-value <0.0001 (****). A.U: arbitrary units.

immune functional nodes were related to pathological response to neoadjuvant chemotherapy. This correlation did not depend on the molecular subtype, indicating that a Systems Biology approach complements knowledge obtained from other research methods. Non-directed probabilistic graphical models allow managing large gene datasets and underscoring lots of gene interactions, many of which have not been previously described.

The activity of immune nodes was higher in tumors attaining a CR and decreased with the intensity of response. Tumors progressing on chemotherapy also showed increased activity in the nodes "Cell division 1" (node 17) and "Cell division 2" (node 18). These results suggest that the patient's immune system plays a crucial role in the response to neoadjuvant chemotherapy. Previous studies suggest that conventional therapies are effective in patients exhibiting some degree of immune activation in the tumor [16], supporting our findings. Chemotherapy may mediate the "immunogenic" death of tumor cells, thus facilitating an immune response against the disease [17].

As expected, all tumors attaining a CR- regardless of molecular subtype- showed significantly higher levels of cell lineage markers (CD4, CD8 and CD20) as well as immune-activating (IGKC, CXCL9, CCL5, CXCL13) and immunosuppressive markers (IDO1, PD1), suggesting a greater infiltration of immune cells High tumor-infiltrating lymphocytes (TILs) levels have been associated with increased CR rates in ER+ HER2+/- tumors [18] and also in TNBC [19]. However, high levels of PD-1+ TILs or Foxp3+ TILs have been related with poor prognosis [18]. Therefore, immune cell subpopulation profiles could help to predict response to neoadjuvant chemotherapy.

Basal-like and HER2+ subgroups have been associated with highest CR rates as opposed to Luminal and Normal-like tumors. However, the genes that were associated with CR in Basal-like subgroup were not associated with CR in the HER2+ subgroup, suggesting that different sets of genes are associated with CR in the different molecular subtypes [20, 21]. In the present cohort, Basal-like tumors had the highest CR rate, as expected. However, the CR rate was poor in HER2+ tumors, probably because these patients did not receive anti-HER2 targeted therapy. On the other hand, BL1 tumors achieved a CR more commonly than other TNBC subtypes, as previously described [22]. Although node "Immune response (MHCII)" (node 9) had higher activity in Luminal A and Normal subtypes, the remaining nodes related to immune response showed increased activity in Basal-like tumors. This agrees with the fact that, in general, there are far fewer TILs in luminal disease than in HER2 or TNBC subtypes [23]. In fact, even though increased TILs concentrations are associated with increased frequency of response to neoadjuvant chemotherapy in all breast cancer subtypes, there is a different effect of TILs on survival in TNBC and luminal breast cancer. Increased TILs concentrations are associated with longer survival in TNBC and HER2+ disease, but not in luminal-HER2-negative tumours [24], suggesting again a differences in the biology of the immunological infiltrate across molecular subtypes.

One possible explanation of the higher "Immune response (MHCII)" (node 9) activity in Luminal A subtype could be the contribution of different immune cell types. Most types of immune cells were increased in TNBC compared with luminal-HER2- negative breast cancer. In TNBC, the presence of many immune cell subtypes, including B cells, T cells, and macrophages, were linked to improved survival [24]. By contrast, in luminal-HER2-negative breast cancer, the presence of T cells was not prognostic for survival and the only cell types linked to improved prognosis were B and myeloid dendritic cells [24], which are MHCII presenting cells. Taking this into account, it would be interesting to perform further studies about MHCII presenting cells infiltration in luminal subtypes.

On the other hand, Basal-like tumors also had the highest activity in the node "Cell cycle" (nodes 17 and 18), which is in accordance with the fact proliferation renders tumor cells more sensitive to chemotherapy [6].

The neoadjuvant setting is appealing in the field of drug development because it allows early evaluation of efficacy. However, not all patients benefit from this approach, so markers predicting response to neoadjuvant chemotherapy are clearly necessary, as neoadjuvant therapy may have some drawbacks, such as promoting metastasis in some cases [25]. Our results suggest that immune activation in the tumor may identify responders. Although validation is needed, the use of these markers can help to determine the future use of neoadjuvant chemotherapy in breast cancer.

MATERIALS AND METHODS

Patient's and samples origin and characteristics

A breast cancer tumor dataset, including gene expression and clinical data, was obtained from the Gene Expression Omnibus (GSE41998) and from a phase II trial (NCT00455533). Gene expression profiling was measured using an Affymetrix GeneChip, normalized and log2 transformed. Surgical specimens were evaluated by a pathologist at each study site. The pathological response was evaluated as the primary endpoint. A pathological complete response (CR) was defined by no histologic evidence of residual invasive adenocarcinoma in the breast and axillary lymph nodes, with or without the presence of ductal carcinoma in situ [12]. Responses were categorized as complete, partial, stabilization or progression.

Gene expression data preprocessing

The PAM50 method was used as previously described to assign a molecular subtype to each sample [13]. Lehmann subtypes for TNBC were assigned in two steps [14]. First, samples were assigned to Lehmann's seven subtypes using centroids constructed from 77 previously assigned tumors in GSE31519 dataset. Then, the IM and MSL groups were redefined as previously described [14].

Probabilistic graphical models construction

A functional structure was developed using undirected probabilistic graphical models (PGMs) as previously described [10]. Briefly, PGMs compatible with high-dimensionality were chosen. The result is an undirected graphical model with local minimum Bayesian Information Criterion. DAVID 6.8 was used to assign a biological function to each node in the networks, using "homo sapiens" as background list and selecting only GOTERM-FAT and Biocarta and KEGG pathways categories. Functional activity of each node was calculated by the mean expression of the genes in each node. To visualize node activities, data from all tumors used to construct the network were mean centred prior to its inclusion into the network.

Statistics and software suites

Differences between groups were assessed using Kruskal-Wallis test, Mann-Whitney test and Dunn`s multiple comparisons test using GraphPad Prism 6. Box-and-whisker plots are Tukey boxplots. All p-values were two-sided and $p<0.05$ was considered statistically significant. Ordinal logistic regression analysis was performed in SAS using logistic procedure. Network analyses were performed in MeV and Cytoscape 3.2.1 software suites.

Abbreviations

AC: doxorubicin/cyclophosphamide; BL1: Basal-like 1; BL2: Basal-like 2; CR: complete response; ER: Estrogen receptor; HER2: Human epidermal growth factor receptor 2; IM: immunomodulatory; MSL: mesenchymal stem-like; PD: progressive disease; PGMs: probabilistic graphical models; PR: Partial response; PR: Progesterone receptor; SD: stable disease; TNBC: Triple Negative Breast Cancer.

Author contributions

All the authors have directly participated in the preparation of this manuscript and have approved the final version submitted. JMA, HN and PM contributed the probabilistic graphical models. MD-A contributed the GPR rule calculation. AZ-M, LT-F and GP-V performed the statistical analysis, the probabilistic graphical model interpretation and the gene ontology analyses. AZ-M drafted the manuscript. AZ-F, AG-P, JAFV, JF and EE conceived of the study and participated in its design and interpretation. AG-P, JAFV, PZ and EE supported the manuscript drafting. AG-P and JAFV coordinated the study. All the authors have read and approved the final manuscript.

CONFLICTS OF INTEREST

JAFV, EE and AG-P are shareholders in Biomedica Molecular Medicine SL. LT-F is an employee of Biomedica Molecular Medicine SL. The other authors declare no potential conflicts of interest.

FUNDING

This study was supported by Instituto de Salud Carlos III, Spanish Economy and Competitiveness Ministry, Spain and co-funded by the FEDER program, "Una forma de hacer Europa" (PI15/01310). LT-F is supported by the Spanish Economy and Competitiveness Ministry (DI-15-07614). The funders had no role in the study design, data collection and analysis, decision to publish or preparation of the manuscript.

REFERENCES

1. Ferlay J, Soerjomataram I, Dikshit R, Eser S, Mathers C, Rebelo M, Parkin DM, Forman D, Bray F. Cancer incidence and mortality worldwide: sources, methods and major patterns in GLOBOCAN 2012. Int J Cancer. 2015; 136:E359-86.
2. Dawson SJ, Rueda OM, Aparicio S, Caldas C. A new genome-driven integrated classification of breast cancer and its implications. EMBO J. 2013; 32:617-28.
3. Honrado E, Benítez J, Palacios J. The pathology of hereditary breast cancer. Hered Cancer Clin Pract. 2004; 2:131-38.
4. Perou CM, Sørlie T, Eisen MB, van de Rijn M, Jeffrey SS, Rees CA, Pollack JR, Ross DT, Johnsen H, Akslen LA, Fluge O, Pergamenschikov A, Williams C, et al. Molecular portraits of human breast tumours. Nature. 2000; 406:747-52.
5. Kennecke H, Yerushalmi R, Woods R, Cheang MC, Voduc D, Speers CH, Nielsen TO, Gelmon K. Metastatic behavior

of breast cancer subtypes. J Clin Oncol. 2010; 28:3271-77.

6. van de Vijver MJ, He YD, van't Veer LJ, Dai H, Hart AA, Voskuil DW, Schreiber GJ, Peterse JL, Roberts C, Marton MJ, Parrish M, Atsma D, Witteveen A, et al. A gene-expression signature as a predictor of survival in breast cancer. N Engl J Med. 2002; 347:1999-2009.
7. Badve S, Dabbs DJ, Schnitt SJ, Baehner FL, Decker T, Eusebi V, Fox SB, Ichihara S, Jacquemier J, Lakhani SR, Palacios J, Rakha EA, Richardson AL, et al. Basal-like and triple-negative breast cancers: a critical review with an emphasis on the implications for pathologists and oncologists. Mod Pathol. 2011; 24:157-67.
8. Smith IC, Heys SD, Hutcheon AW, Miller ID, Payne S, Gilbert FJ, Ah-See AK, Eremin O, Walker LG, Sarkar TK, Eggleton SP, Ogston KN. Neoadjuvant chemotherapy in breast cancer: significantly enhanced response with docetaxel. J Clin Oncol. 2002; 20:1456-66.
9. Eroles P, Bosch A, Pérez-Fidalgo JA, Lluch A. Molecular biology in breast cancer: intrinsic subtypes and signaling pathways. Cancer Treat Rev. 2012; 38:698-707.
10. Gámez-Pozo A, Berges-Soria J, Arevalillo JM, Nanni P, López-Vacas R, Navarro H, Grossmann J, Castaneda CA, Main P, Díaz-Almirón M, Espinosa E, Ciruelos E, Fresno Vara JÁ. Combined label-free quantitative proteomics and microRNA expression analysis of breast cancer unravel molecular differences with clinical implications. Cancer Res. 2015; 75:2243-53.
11. de Anda-Jáuregui G, Velázquez-Caldelas TE, Espinal-Enríquez J, Hernández-Lemus E. Transcriptional Network Architecture of Breast Cancer Molecular Subtypes. Front Physiol. 2016; 7: 568.
12. Horak CE, Pusztai L, Xing G, Trifan OC, Saura C, Tseng LM, Chan S, Welcher R, Liu D. Biomarker analysis of neoadjuvant doxorubicin/cyclophosphamide followed by ixabepilone or Paclitaxel in early-stage breast cancer. Clin Cancer Res. 2013; 19:1587-95.
13. Parker JS, Mullins M, Cheang MC, Leung S, Voduc D, Vickery T, Davies S, Fauron C, He X, Hu Z, Quackenbush JF, Stijleman IJ, Palazzo J, et al. Supervised risk predictor of breast cancer based on intrinsic subtypes. J Clin Oncol. 2009; 27:1160-67.
14. Lehmann BD, Jovanović B, Chen X, Estrada MV, Johnson KN, Shyr Y, Moses HL, Sanders ME, Pietenpol JA. Refinement of triple-negative breast cancer molecular subtypes: implications for neoadjuvant chemotherapy selection. PLoS One. 2016; 11:e0157368.
15. Gámez-Pozo A, Trilla-Fuertes L, Berges-Soria J, Selevsek N, López-Vacas R, Díaz-Almirón M, Nanni P, Arevalillo JM, Navarro H, Grossmann J, Gayá Moreno F, Gómez Rioja R, Prado-Vázquez G, et al. Functional proteomics outlines the complexity of breast cancer molecular subtypes. Sci Rep. 2017; 7:10100.
16. Denkert C, Loibl S, Noske A, Roller M, Müller BM, Komor M, Budczies J, Darb-Esfahani S, Kronenwett R, Hanusch C, von Törne C, Weichert W, Engels K, et al. Tumor-associated lymphocytes as an independent predictor of response to neoadjuvant chemotherapy in breast cancer. J Clin Oncol. 2010; 28:105-13.
17. Ghiringhelli F, Apetoh L. The interplay between the immune system and chemotherapy: emerging methods for optimizing therapy. Expert Rev Clin Immunol. 2014; 10:19-30.
18. Yu X, Zhang Z, Wang Z, Wu P, Qiu F, Huang J. Prognostic and predictive value of tumor-infiltrating lymphocytes in breast cancer: a systematic review and meta-analysis. Clin Transl Oncol. 2016; 18:497-506.
19. Denkert C, von Minckwitz G, Brase JC, Sinn BV, Gade S, Kronenwett R, Pfitzner BM, Salat C, Loi S, Schmitt WD, Schem C, Fisch K, Darb-Esfahani S, et al. Tumor-infiltrating lymphocytes and response to neoadjuvant chemotherapy with or without carboplatin in human epidermal growth factor receptor 2-positive and triple-negative primary breast cancers. J Clin Oncol. 2015; 33:983-91.
20. Kuerer HM, Newman LA, Smith TL, Ames FC, Hunt KK, Dhingra K, Theriault RL, Singh G, Binkley SM, Sneige N, Buchholz TA, Ross MI, McNeese MD, et al. Clinical course of breast cancer patients with complete pathologic primary tumor and axillary lymph node response to doxorubicin-based neoadjuvant chemotherapy. J Clin Oncol. 1999; 17:460-69.
21. Rouzier R, Perou CM, Symmans WF, Ibrahim N, Cristofanilli M, Anderson K, Hess KR, Stec J, Ayers M, Wagner P, Morandi P, Fan C, Rabiul I, et al. Breast cancer molecular subtypes respond differently to preoperative chemotherapy. Clin Cancer Res. 2005; 11:5678-85.
22. Masuda H, Baggerly KA, Wang Y, Zhang Y, Gonzalez-Angulo AM, Meric-Bernstam F, Valero V, Lehmann BD, Pietenpol JA, Hortobagyi GN, Symmans WF, Ueno NT. Differential response to neoadjuvant chemotherapy among 7 triple-negative breast cancer molecular subtypes. Clin Cancer Res. 2013; 19:5533-40.
23. Savas P, Salgado R, Denkert C, Sotiriou C, Darcy PK, Smyth MJ, Loi S. Clinical relevance of host immunity in breast cancer: from TILs to the clinic. Nat Rev Clin Oncol. 2016; 13:228-41.
24. Denkert C, von Minckwitz G, Darb-Esfahani S, Lederer B, Heppner BI, Weber KE, Budczies J, Huober J, Klauschen F, Furlanetto J, Schmitt WD, Blohmer JU, Karn T, et al. Tumour-infiltrating lymphocytes and prognosis in different subtypes of breast cancer: a pooled analysis of 3771 patients treated with neoadjuvant therapy. Lancet Oncol. 2018; 19:40-50.
25. Karagiannis GS, Pastoriza JM, Wang Y, Harney AS, Entenberg D, Pignatelli J, Sharma VP, Xue EA, Cheng E, D'Alfonso TM, Jones JG, Anampa J, Rohan TE, et al. Neoadjuvant chemotherapy induces breast cancer metastasis through a TMEM-mediated mechanism. Sci Transl Med. 2017; 9.

SCIENTIFIC REPORTS

OPEN

A novel approach to triple-negative breast cancer molecular classification reveals a luminal immune-positive subgroup with good prognoses

Received: 11 June 2018
Accepted: 19 December 2018
Published: xx xx xxxx

Guillermo Prado-Vázquez[1,2], Angelo Gámez-Pozo[1,2], Lucía Trilla-Fuertes[2], Jorge M. Arevalillo[4], Andrea Zapater-Moros[1,2], María Ferrer-Gómez[1], Mariana Díaz-Almirón[3], Rocío López-Vacas[1], Hilario Navarro[4], Paloma Main[5], Jaime Feliú[6,7], Pilar Zamora[6], Enrique Espinosa[6,7] & Juan Ángel Fresno Vara[1,7]

Triple-negative breast cancer is a heterogeneous disease characterized by a lack of hormonal receptors and HER2 overexpression. It is the only breast cancer subgroup that does not benefit from targeted therapies, and its prognosis is poor. Several studies have developed specific molecular classifications for triple-negative breast cancer. However, these molecular subtypes have had little impact in the clinical setting. Gene expression data and clinical information from 494 triple-negative breast tumors were obtained from public databases. First, a probabilistic graphical model approach to associate gene expression profiles was performed. Then, sparse k-means was used to establish a new molecular classification. Results were then verified in a second database including 153 triple-negative breast tumors treated with neoadjuvant chemotherapy. Clinical and gene expression data from 494 triple-negative breast tumors were analyzed. Tumors in the dataset were divided into four subgroups (luminal-androgen receptor expressing, basal, claudin-low and claudin-high), using the cancer stem cell hypothesis as reference. These four subgroups were defined and characterized through hierarchical clustering and probabilistic graphical models and compared with previously defined classifications. In addition, two subgroups related to immune activity were defined. This immune activity showed prognostic value in the whole cohort and in the luminal subgroup. The claudin-high subgroup showed poor response to neoadjuvant chemotherapy. Through a novel analytical approach we proved that there are at least two independent sources of biological information: cellular and immune. Thus, we developed two different and overlapping triple-negative breast cancer classifications and showed that the luminal immune-positive subgroup had better prognoses than the luminal immune-negative. Finally, this work paves the way for using the defined classifications as predictive features in the neoadjuvant scenario.

Breast cancer (BC) causes 450,000 deaths every year worldwide[1]. BC is clinically and genetically heterogeneous[2], and this heterogeneity has led to subdivisions in an attempt to treat patients more efficiently. The classical

[1]Molecular Oncology & Pathology Lab, INGEMM, La Paz University Hospital Health Research Institute-IdiPAZ, Madrid, Spain. [2]BioMedica Molecular Medicine SL, La Paz University Hospital Health Research Institute-IdiPAZ, Madrid, Spain. [3]Biostatistics Unit, La Paz University Hospital Health Research Institute-IdiPAZ, Madrid, Spain. [4]Department of Statistics, Operational Research and Numerical Analysis, National University of Distance Education (UNED), Madrid, Spain. [5]Department of Statistics and Operations Research, Faculty of Mathematics, Complutense University of Madrid, Madrid, Spain. [6]Medical Oncology Service, La Paz University Hospital Health Research Institute-IdiPAZ, Madrid, Spain. [7]CIBERONC, La Paz University Hospital Health Research Institute-IdiPAZ, Madrid, Spain. Correspondence and requests for materials should be addressed to J.Á.F.V. (email: juanangel.fresno@salud.madrid.org)

Q1 Q2 Q3 Q4 Q5 Q6

categorization considers the expression of hormonal receptors (estrogen receptors [ERs], and progesterone receptors [PRs]) and human epidermal growth factor receptor 2 (HER2) expression, because this determines the possibility of treatment with hormones and anti-HER2 therapies, respectively.

Triple-negative breast cancer (TNBC) is defined by a lack of ER and PR expression and a lack of HER2 overexpression. TNBC comprises a heterogeneous group of tumors. In 2000, Perou *et al*. proposed a classification of BC based on gene expression patterns. Most triple-negative tumors are included in the so-called basal-like molecular subgroup[3], although both categories have up to 30% discordance[4].

Several studies have developed specific molecular classifications for TNBC. For example, Rody *et al*. defined metagenes that distinguished molecular subsets within the group[5]. Lehmann *et al*. identified seven molecular subgroups: unstable; basal-like 1; basal-like 2; immunomodulatory; mesenchymal (MES)-like; mesenchymal stem-like (MSL); and luminal androgen receptor (LAR)[6]. The Immunomodulatory and MSL subtypes have recently been refined[7]. Burstein *et al*. applied non-negative matrix factorization and defined four subgroups: basal-like immune active; basal-like immune suppressed; mesenchymal; and luminal AR[8]. Other classifications have also been proposed by Sabatier[9], Prat[10], Jézéquel[11], and Milioli[12]. Despite these extensive studies, the designation of TNBC molecular subtypes has had little impact in the clinical setting.

The so-called cancer stem cell hypothesis could provide a different way to categorize BC. It theorizes that cancer derives from a stem cell compartment that undergoes an abnormal and poorly regulated process of organogenesis analogous to many aspects of normal stem cells[13–15]. Depending on the activation point of these cancer stem cells, tumors will have varying characteristics. Poorly differentiated breast tumors would arise from the most primitive stem cells[14]. This hypothesis contextualizes BC molecular groups[1] in a development framework. Moreover, molecular characterization of the claudin (CLDN)-low subtype reveals that these tumors are significantly enriched in epithelial-mesenchymal transition and stem cell-like features, while showing a low expression of luminal and proliferation-associated genes[16].

In the present study, we applied probabilistic graphical models to a previously published TNBC cohort[5]. This technique allows exploring the molecular information from a functional perspective. Our aim was to tackle the molecular analysis of TNBC from a broad perspective, such as the cancer stem cell hypothesis, to provide a classification with clearer clinical implications.

Methods

TNBC gene expression and clinical data. Gene expression data from TNBC tumors and available clinical follow-up information were obtained from GSE31519. Gene expression values were magnitude normalized, and then $\log_2$ was calculated. The *Limma* R package[17] was applied to avoid the batch effect. Finally, the complete dataset was mean centered. The probe with the highest variance of each gene within all patients was selected. The results obtained with the first database were then applied to a second database of patients treated with neoadjuvant chemotherapy, GSE25066. GSE25066 data was magnitude normalized and $\log_2$ was calculated just as with GSE31519.

Probabilistic graphical model analysis. A probabilistic graphical model compatible with a high-dimensionality approach to associate gene expression profiles, including the most variable 2000 genes, was performed as previously described[18]. Briefly, the resulting network, in which each node represents an individual gene, was split into several branches to identify functional structures within the network. Then, we used gene ontology analyses to investigate which function or functions were overrepresented in each branch, using the functional annotation chart tool provided by DAVID 6.8 beta[19]. We used "homo sapiens" as a background list and selected only GOTERM-DIRECT gene ontology categories and Biocarta and KEGG pathways. Functional nodes were composed of nodes presenting a gene ontology enriched category. To measure the functional activity of each functional node, the mean expression of all the genes included in one branch related to a concrete function was calculated. Differences in functional node activity were assessed by class comparison analyses. Finally, metanodes were defined as groups of related functional nodes using nonsupervised hierarchical clustering analyses.

Sparse k-means classification. Sparse k-means was used to establish the optimal number of tumor groups. This method uses the genes included in each node and metanode, as previously described[20]. Briefly, classification consistency was tested using random forest. An analysis using the consensus clustering algorithm[21] as applied to the data containing the variables that were selected by the sparse K-means method[22] has provided an optimum classification into two subtypes in previous studies[20]. In order to transfer the newly defined classification from the main dataset to other datasets, we constructed centroids for each defined subgroup, using genes included in various metanodes.

Assignation to groups defined by other molecular classifications. Tumors in the main dataset were assigned to a single group according to previously defined molecular classifications: PAM50 + CLDN low was assigned using the single sample predictor[10]. Burstein's four subtypes were assigned using an 80-gene signature[8]. The TNBC4 type was performed in two steps: first, Lehmann's seven subtypes were assigned using centroids constructed from 77 tumors included in the dataset that was previously assigned, and then Immunomodulatory and MSL groups were redefined as previously described[7].

Statistical analyses and software suites. Survival curves were estimated using Kaplan–Meier analyses and compared with the log-rank test, using relapse free survival (RFS) as the end point. RFS was defined as the time between the day of surgery and the date of distant relapse or last date of follow-up. Correlations were assessed using Pearson's r and linear regression. Differences in functional node activity between groups were assessed by the Kruskal–Wallis test, and multiple comparisons were assessed using the Dunn's multiple comparisons test. Box-and-whisker plots are Tukey boxplots. All p-values were two-sided, and $P < 0.05$ was considered

	Main Dataset	Neoadjuvant dataset	p-value
Number of patients	494	153	
Tumor Size			
T1	99 (20%)	9 (6%)	<0.0001
>T1	276 (56%)	144 (94%)	
NA	119 (24%)		
Tumor Grade			
G1&2	103 (21%)	16 (10%)	0.0001
G3	280 (57%)	124 (81%)	
NA	111 (22%)		
Lymph node status			
N0	251 (51%)	37 (24%)	<0.0001
N1	68 (14%)	116 (76%)	
NA	175 (35%)		
Adjuvant Chemotherapy			
No	257 (52%)		
Yes	71 (14%)		
NA	166 (34%)		
Pathological Response			
RD		95 (62%)	
pCR		53 (34%)	

Table 1. Clinical features of the main and neoadjuvant datasets. Size data is divided into T1 (<2 cm) and >T1 (>2 cm) tumors; grade is classified as G1&2 (well or moderately differentiated tumors) or G3 (poorly differentiated tumors); lymph node status represents lymph node invasion (N0: no invasion; N1: invasion or metastasis); and the adjuvant chemotherapy column comprises patients who had been treated with adjuvant chemotherapy or not. The pathological response column stands for the response to neoadjuvant treatment (RD: residual disease; pCR: pathological complete response). The chi-squared test confirmed that both cohorts are different regarding clinical parameters and treatment.

statistically significant. Expression data and network analyses were performed in MeV and Cytoscape software suites[23]. The SPSS v16 software package, GraphPad Prism 5.0 and R v2.15.2 (with the Design software package 0.2.3) were used for all the statistical analyses.

Results

Gene expression and clinical data. Gene expression data and clinical information from 579 TNBC tumors were obtained from GSE31519. Some 85 samples were excluded because the patients had been treated with neoadjuvant chemotherapy or a different platform had been used. As a consequence, the data from 494 TNBC tumors from GSE31519 were used in subsequent analyses. Gene expression was normalized, the batch effect was corrected and the most variant probe was selected for each gene. The resulting dataset, including expression values from 13,146 genes will be referred to as the main dataset from now on.

Gene expression data from 508 breast cancer samples treated with neoadjuvant taxane-anthracycline chemotherapy were retrieved from GSE25066. A total 153 of these 508 samples were identified as TNBC.

Clinical features. All available clinical features of the main dataset and the neoadjuvant dataset are presented in Table 1. The main dataset's population of tumors tended to be large (>T1 in 56% of the population), poorly differentiated (G3 in 57% of the samples), with no node invasion (N0 in 51% of the samples) and most of the patients were not treated with adjuvant chemotherapy (52%). The neoadjuvant dataset's population of tumors tended to be T2 (44%) and T3 (32%), poorly differentiated (G3 in 81% of the samples), and N1 (46%) with 32% of the patients achieving a complete pathological response after neoadjuvant treatment.

Molecular characterization of TNBC. A gene expression-based network, including the 2000 most variant genes in the development dataset, was constructed using a probabilistic graphical model (PGM) (Fig. 1). The functional structure of the network was explored using gene ontology analyses, and 26 functional nodes were defined (Fig. 1 and Sup. File 1). Functional node activity was calculated and relationships between nodes were assessed using a hierarchical clustering (HCL) analysis (Sup. File 2). Functional node 1 is composed of 34 genes, including the CLDN3, CLDN4 and CLDN7 genes. On the other hand, functional nodes 15 (chemokine activity), 16 (major histocompatibility complex class II receptor activity), 17 (immune response) and 18 (antigen binding) were related to various aspects of the immune response and clustered together as an "immune metanode" in the HCL analysis (Sup. File 2). Additionally, functional node 19 contained genes related to the peroxisome proliferator-activated receptor (PPAR) signaling pathway, and functional node 24 contained genes involved in the G1/S transition of mitotic cell cycle (Sup. File 1).

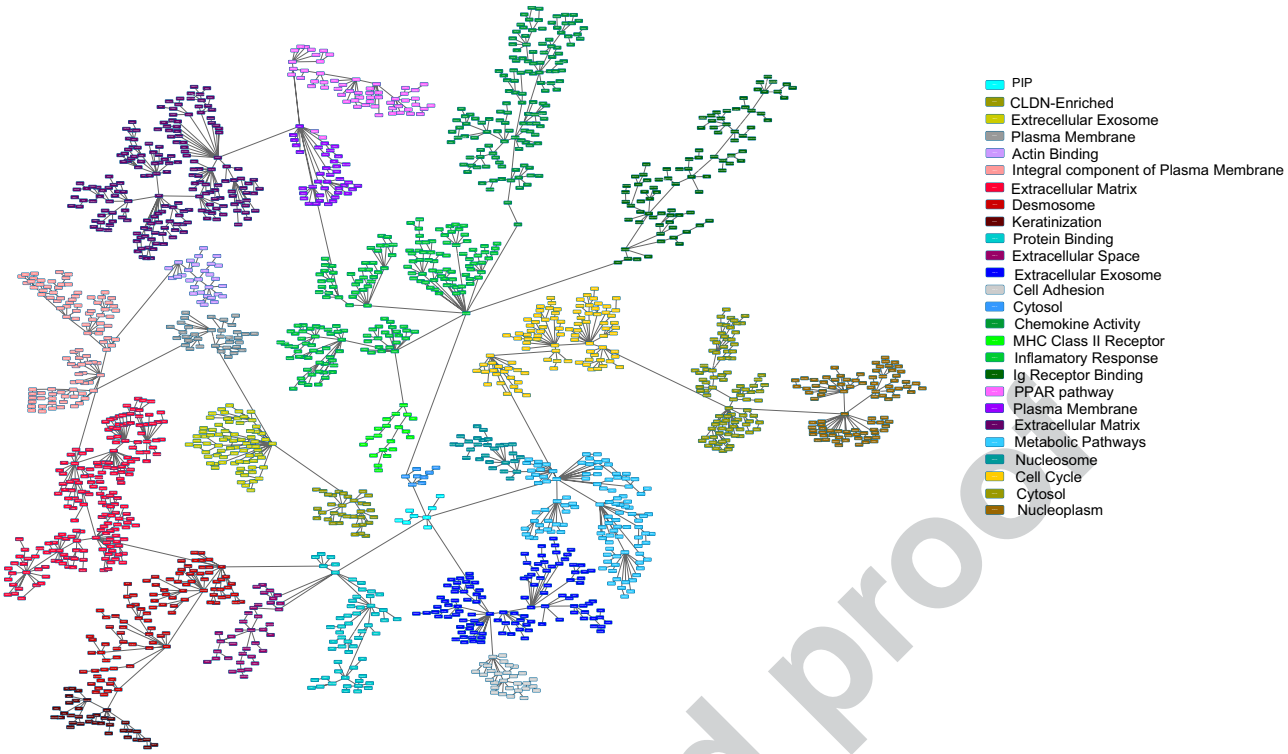


Figure 1. PGM resulting network; each functional node is encoded from 0 to 26. Each box (node) represents one gene, and lines (edges) connect genes with related expression. Functional nodes are represented by the same color, and metanodes are presented the same color palette, with basal nodes in red, luminal nodes in blue and immune nodes in green.

We then used the method described by Rody *et al*. to assess 15 metagenes (series of genes known to be related to one specific biological function or characteristic)[5]. Genes within a given metagene appeared close to each other in our network. Additionally, related metagenes, i.e., B-cell and IL-8 metagenes, also appeared close to each other (Fig. 2).

Functional nodes 5, 6, 7, 8 and 10 in our network had different gene ontologies related to an integral component of the plasma membrane, extracellular matrix, desmosomes, keratinization, and extracellular space, respectively. However, these five nodes appeared to correlate in the HCL analysis (Sup. File 2) and included genes from Rody's basal-like metagene (Fig. 2). Thus, from now on, these five functional nodes were grouped as the basal metanode (Fig. 1). In the same way, functional nodes 0, 9, 11, 14, 22 and 23 were related to protein binding, extracellular exosomes, sequence-specific DNA binding, metabolic pathways and nucleosomes, respectively, again grouped together in the HCL analysis and including genes from Rody's apocrine/luminal metagene, so they were defined as the luminal metanode (Fig. 1).

Cellular classification. The sparse k-means method was used to group samples into a limited number of clusters based on functional nodes and metanodes. Samples from the basal and luminal metanodes and the CLDN-enriched functional node were each divided into two groups. Mimicking the cancer stem cell hypothesis, we established the following workflow (Fig. 3): Samples with high luminal metanode activity were classified as the luminal androgen receptor group (LAR). Tumors showing low luminal metanode activity and high basal metanode activity from the basal subgroup were classified as basal. Finally, tumors with low activity in both the basal and luminal metanodes were screened for CLDN-enriched node expression. Samples showing low activity for the CLDN-enriched functional node were categorized as CLDN-low, whereas samples showing high activity for CLDN-enriched functional node were labeled as CLDN-high (Fig. 3).

From the 494 samples in the main dataset, the cellular classification defined 91 (18%) LAR, 53 (11%) CLDN-low, 310 (63%) basal and 40 (8%) CLDN-high samples. Only 7 (1.5%) samples showed high activity in both the luminal and basal metanodes (Table 2).

Clinical characteristics from the various entities of cellular classification are shown in Table 3. Basal subtype tumors were mostly small-sized, poorly differentiated and without lymph node infiltration. The CLDN-high subtype tumors were large, had poor differentiation and no lymph node infiltration. The CLDN-low as well as the LAR tumors were large, more differentiated and showed more infiltration than the basal and CLDN-high tumors. Cellular classification does not show a significant relationship to RFS (Sup. File 3), nor did basal and luminal metanode activities show prognostic value. CLDN-high tumors showed a trend toward a poorer prognosis than CLDN-low, but again, the differences were not significant.

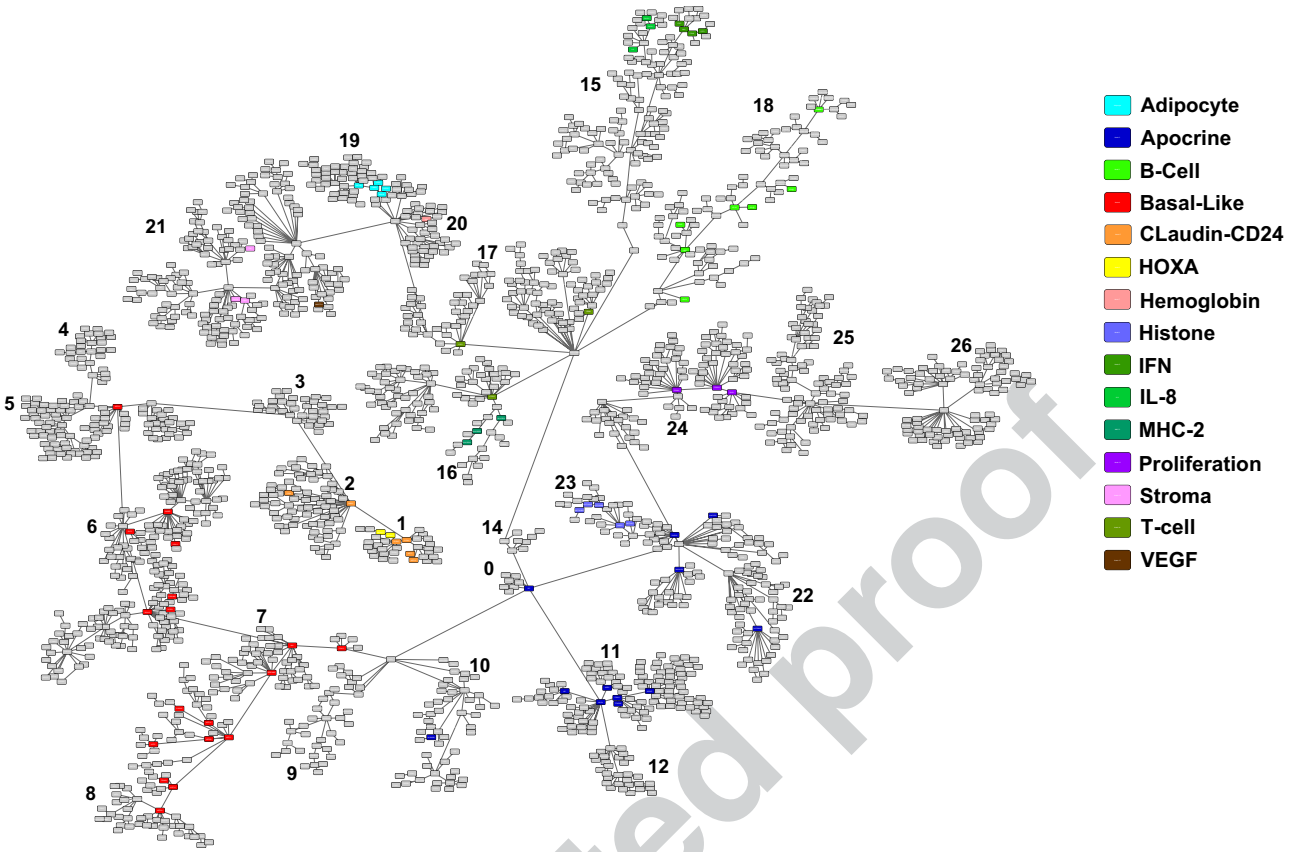


Figure 2. PGM represents the resulting network in which each functional node is encoded from 0 to 26, each box (node) represents one gene and lines (edges) connect genes with related expression. Genes from Rody's metagenes are represented by different colors.

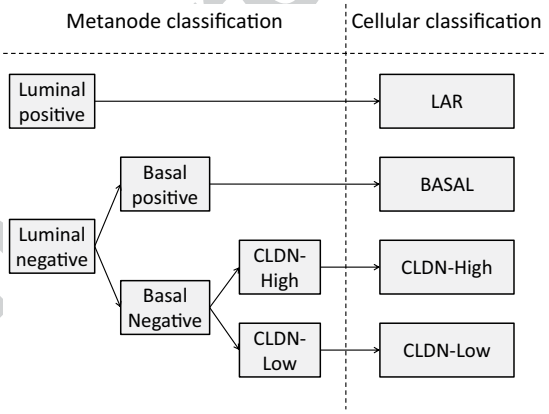


Figure 3. Workflow from the sparse k-means groups in each metanode to the final cellular classification.

Activity of functional nodes in cellular groups. The activity of the main functional nodes was assessed in each cellular group. CLDN-low tumors had lower activity than every other tumor subgroup in the functional nodes related to alpha-amylase activity and regulation of actin cytoskeleton, and higher activity than the other subgroups in the haptoglobin binding functional node. CLDN-high tumors had lower activity than basal tumors in the actin binding functional node, higher activity than tumors belonging to any other subgroup in chemokine activity functional node and lower activity than CLDN-low and LAR subtypes in the haptoglobin binding functional node. Basal tumors had higher activity than any other tumor in the functional nodes related to cell adhesion and regulation of the actin cytoskeleton. Finally, LAR tumors had lower activity in the nodes related to cell adhesion, G1/S transition of mitotic cell cycle and chemokine activity (Sup. File 4).

Luminal	N	Basal	N	CLDN	Tumors	% of total	Cellular	N
—	403 (82%)	—	93 (23%)	High	40 (43%)	8%	CLDN-High	40 (8%)
				Low	53 (57%)	11%	CLDN-Low	53 (11%)
		+	310 (77%)	High	245 (79%)	50%	Basal	310 (63%)
				Low	65 (21%)	13%		
+	91 (18%)	—	84 (92%)	High	79 (94%)	16%	LAR	91 (18%)
				Low	5 (6%)	1%		
		+	7 (8%)	High	7 (100%)	1%		
				Low	0	0%		

Table 2. Number of tumors classified in each metanode sparse k-means group and in the cellular classification.

Cellular Classification	Tumor size			Grade			Nodal		
	T1	>T1	p-value	G1 or G2	G3	p-value	N0	N1	p-value
Basal	76 (32%)	163 (68%)	0.169	45 (18%)	199 (82%)	0.015	168 (83%)	35 (17%)	0.262
CLDN-High	2 (7%)	27 (93%)	0.023	5 (14%)	31 (86%)	0.110	19 (83%)	4 (17%)	0.795
CLDN-Low	10 (24%)	32 (76%)	0.853	20 (49%)	21 (51%)	0.005	28 (72%)	11 (28%)	0.313
LAR	11 (17%)	54 (83%)	0.121	33 (53%)	29 (47%)	<0.001	36 (67%)	18 (33%)	0.056
Total	99 (26%)	276 (74%)	—	103 (27%)	280 (73%)	—	251 (79%)	68 (21%)	—

Table 3. Number of tumors with clinical characteristics. T1: tumor smaller than 2 cm; >T1: tumor larger than 2 cm; G3: grade 3; G1 or G2: grade 1 or grade 2; Nodal (N0): no node infiltration; N1: node infiltration. % is calculated using the total amount of a row for each clinical characteristic. Fisher exact test were performed between each group of the cellular classification and the total population (significant p-value = 0.05).

IM negative			IM positive		
Cellular Classification	Tumors	%	Cellular Classification	Tumors	%
Basal	159	68%	Basal	151	58%
CLDN-Low	23	10%	CLDN-Low	30	12%
LAR	42	18%	LAR	49	19%
CLDN-High	11	5%	CLDN-High	29	11%

Table 4. Immune characteristic interaction with cellular classification. According to the chi-squared test, IM characteristics and cellular classification are dependent.

Immune metanode activity: Immune characteristics. On the other hand, taking the immune metanode into account, tumors were split according to their immune (IM) activity. High/low immune activity was defined with the sparse K-means method using genes included in the IM metanode. Some 259 (52%) samples were included in the IM-positive (IM+) group and 235 (48%) were included in the IM-negative (IM−) group (Table 4).

IM+ tumors had a better prognosis than IM- tumors (hazard ratio [HR], 0.7286; 95% confidence interval [CI] 0.5329–0.9961; $P < 0.05$) (Fig. 4A). In addition, the immune metanode activity had a prognostic impact on the groups defined by the cellular classification. Patients with IM+/LAR subtype tumors had a better prognosis than those with IM−/LAR tumors (HR, 0.3474; 95% CI 0.1657–0.7284; $P < 0.05$). Also, patients with IM+/CLDN-high tumors had a better prognosis than those with IM−/CLDN−, although these differences did not reach statistical significance (HR, 0.3556; 95% CI 0.04115–0.9828; $P = 0.057$). IM activity had no impact on the prognosis of the basal and CLDN-low subtypes (Fig. 4B).

Comparison between Cellular classification and PAM50, TNBC4-type and Burstein's classifications. Cellular classification and previous classifications were compared (Fig. 5). The basal subtype is highly enriched in basal-like immune suppressed (BLIS) and basal-like immune associated (BLIA) (Burstein 2015), basal (PAM50 + CLDN-low) and M (Lehmann 2016) subtypes, and it is poorly represented in the LAR subtypes from the Burstein and Lehmann classifications. The CLDN-high subtype is highly enriched in BLIA (Burstein 2015) and BL2 (Lehmann 2016). The CLDN-low subtype is highly enriched in MES (Burstein 2015), LumA (PAM50 + CLDN-low) and BL2 (Lehmann 2016). The LAR subtype is highly enriched in LAR (Burstein 2015), LumA (PAM50 + CLDN-low) and LAR (Lehmann 2016). The LAR subtype is not present in Basal (PAM50) and BL1 (Lehmann) assignations (Fig. 5 and Table 5).

Immune characteristics and previous classifications. The Mesenchymal subtype from the TNBC4 type[7] was highly enriched in IM- samples (148 samples of 187, 80% of all M subtype samples). Also, BL2 was

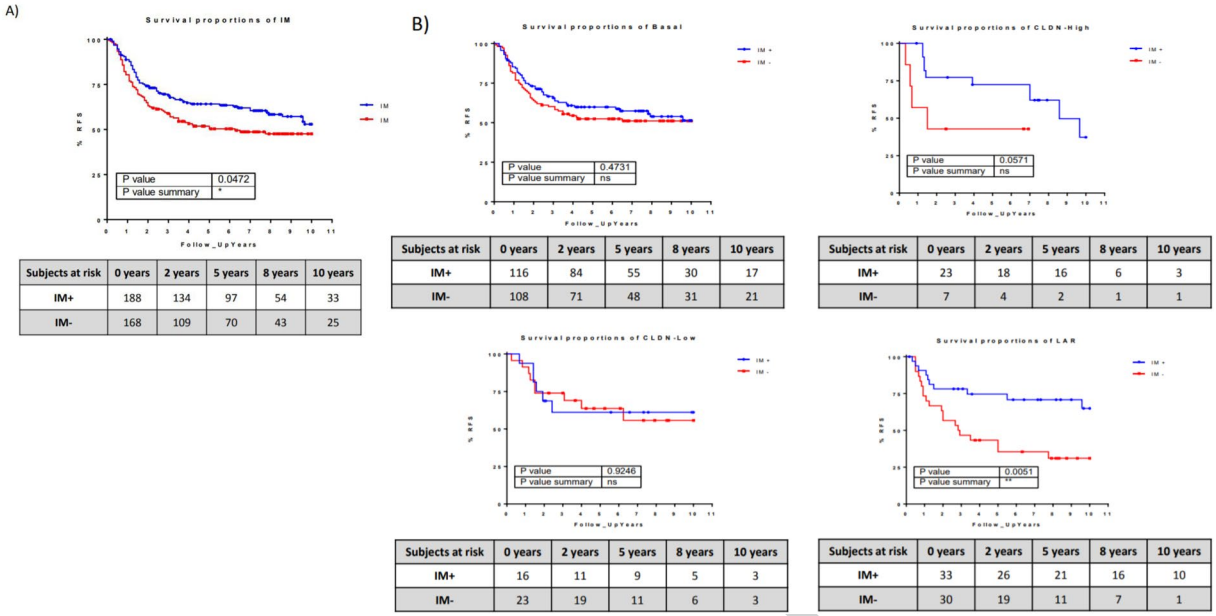


Figure 4. Kaplan-Meier survival curves represent the survival rate of immune-positive and immune-negative tumors in the whole cohort (**A**) and in the four cellular subgroups (**B**).

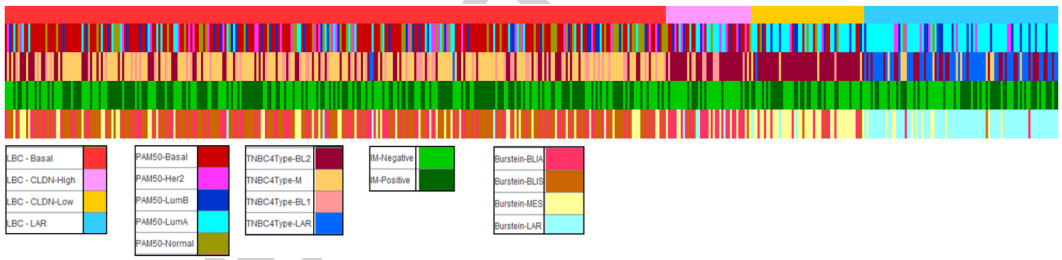
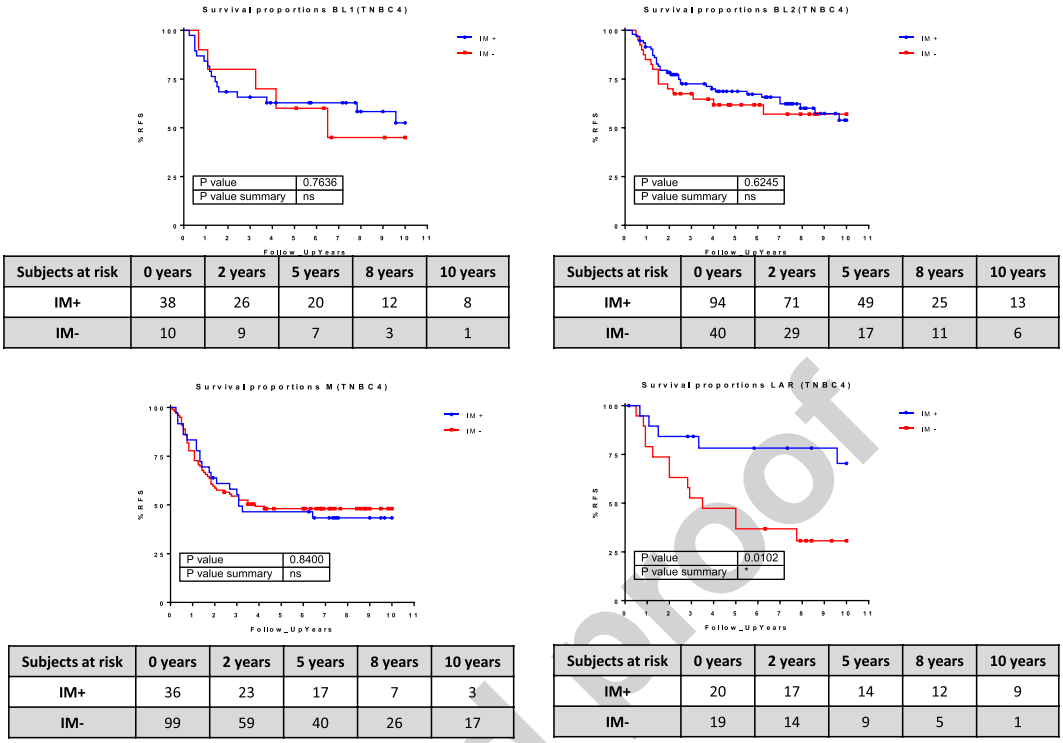


Figure 5. Various molecular classifications compared with the cellular classification. From top to bottom, cellular, PAM50 + CLDN-low, Lehmann 2016 TNBC4 type, immune and Burstein's classifications are presented.

Basal			**CLDN-High**			**CLDN-Low**			**LAR**		
Burstein	**N**	**%**	**Burstein**	**N**	**%**	**Burstein**	**N**	**%**	**Burstein**	**N**	**%**
BLIA	104	34%	BLIA	23	58%	BLIA	11	21%	BLIA	2	2%
BLIS	149	48%	BLIS	3	1%	BLIS	3	6%	BLIS	1	1%
LAR	4	1%	LAR	4	1%	LAR	3	6%	LAR	76	84%
MES	53	17%	MES	10	25%	MES	36	68%	MES	12	13%
PAM50 + CLDN−Low	**N**	**%**	**PAM50 + CLDN−Low**	**N**	**%**	**PAM50 + CLDN−Low**	**N**	**%**	**PAM50 + CLDN−Low**	**N**	**%**
Basal	125	40%	Basal	13	33%	Basal	5	9%	Basal	0	0%
CLDN-Low	76	25%	CLDN-Low	9	23%	CLDN-Low	44	83%	CLDN-Low	13	14%
Her2	23	7%	Her2	6	15%	Her2	1	2%	Her2	8	9%
LumA	25	8%	LumA	7	18%	LumA	1	2%	LumA	52	57%
LumB	27	9%	LumB	4	10%	LumB	4	4%	LumB	16	18%
Normal	34	11%	Normal	1	3%	Normal	0	0%	Normal	2	2%
TNBC4 type	**N**	**%**	**TNBC4 type**	**N**	**%**	**TNBC4 type**	**N**	**%**	**TNBC4 type**	**N**	**%**
BL1	57	18%	BL1	8	20%	BL1	1	2%	BL1	0	0%
BL2	81	26%	BL2	29	73%	BL2	47	89%	BL2	28	31%
LAR	3	1%	LAR	0	0%	LAR	1	2%	LAR	52	57%
M	169	55%	M	3	8%	M	4	8%	M	11	12%

Table 5. Shows comparisons between Cellular classification and PAM50, Lehmann's and Burstein's classifications.



Subjects at risk	0 years	2 years	5 years	8 years	10 years
IM+	38	26	20	12	8
IM-	10	9	7	3	1

Subjects at risk	0 years	2 years	5 years	8 years	10 years
IM+	94	71	49	25	13
IM-	40	29	17	11	6

Subjects at risk	0 years	2 years	5 years	8 years	10 years
IM+	36	23	17	7	3
IM-	99	59	40	26	17

Subjects at risk	0 years	2 years	5 years	8 years	10 years
IM+	20	17	14	12	9
IM-	19	14	9	5	1

Figure 6. Kaplan-Meier survival curves represent the survival rate of immune-positive and immune-negative tumors in the TNBC4-type subgroups.

PAM50 + CLND-low	IM−	IM+
Basal	69 (48%)	74 (52%)
CLDN-low	62 (44%)	80 (56%)
Her2	10 (26%)	28 (74%)
LumA	43 (51%)	42 (49%)
LumB	27 (55%)	22 (57%)
Normal	24 (65%)	13 (35%)

Table 6. Shows immune characteristics in the PAM50+CLDN-low subgroups.

enriched in IM+ samples (135 samples of 185, 72% of all BL2 subtype samples). The IM+ and IM- groups showed no prognostic value for the BL1, BL2 and M groups (Fig. 6). However, patients with IM+ tumors had better prognosis than those with IM− in the LAR group (HR, 0.2896; 95% CI 0.1125–0.7273; $P < 0.05$).

The IM+ and IM- subgroups were evenly distributed in the subtypes defined by PAM50 and CLDN-low, with the exception of the HER2 subtype, which was enriched in IM+ (Table 6).

LumA immune-positive tumors had a better prognosis than immune-negative tumors (HR, 2.638; 95% CI 1.098–6.341; $P < 0.05$). Basal Immune and normal-like immune-positive tumors also showed a trend toward a better prognosis than immunenegative, but the differences were not statistically significant. Finally CLDN-low, LumB and HER2 tumors showed no differences in prognosis related to their immune status (Fig. 7).

Finally, the Burstein subtype BLIA was highly enriched in the IM+ (106 samples of 140, 75%) and the BLIS was highly enriched in the IM- tumors (119 samples of 156, 76%).

Immune-positive and immune-negative tumors had different outcomes in each of the Burstein's subgroups. BLIA, BLIS and LAR immune-positive tumors as well as MES immune-negative tumors had a better prognosis, although the differences were not statistically significant (Fig. 8).

Implications of the cellular classification and the immune characteristic in response to neoadjuvant treatment. Cellular classification was transferred using genes from the basal and luminal metanodes and the CLDN-enriched functional node. Of 153 triple-negative breast cancer tumors, 79 were assigned to the basal subgroup (51%), 8 were assigned to the CLDN-high subgroup (5%), 19 were assigned to the CLDN-low subgroup (12%) and 47 were assigned to the LAR subgroup (31%). The immune characteristic was transferred using genes from the immune metanode. Some 80 samples were immune-negative (52%) and 73 samples were assigned to the immune-positive subgroup (47%) (Table 7).

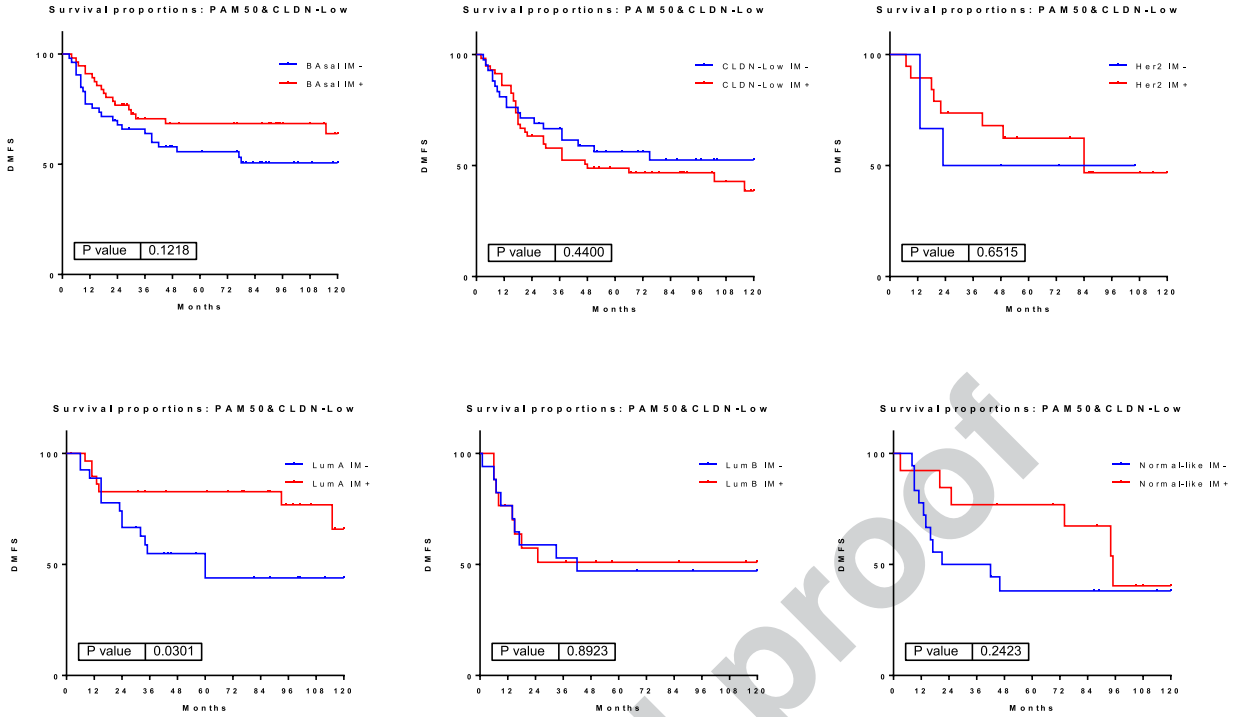
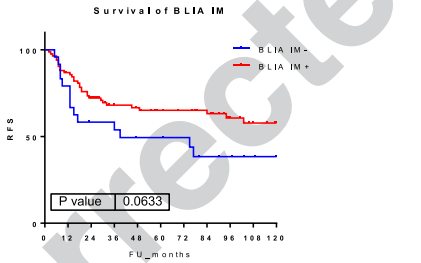
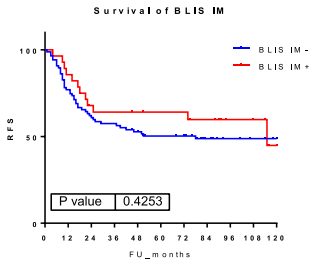


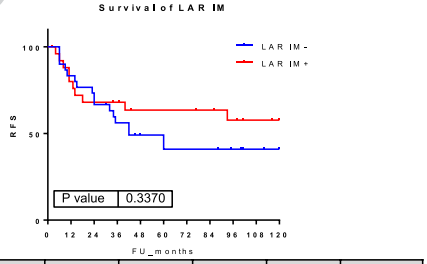
Figure 7. Kaplan-Meier survival curves represent the survival rate of immune-positive and immune-negative tumors in the PAM50 + CLDN-low subgroups.



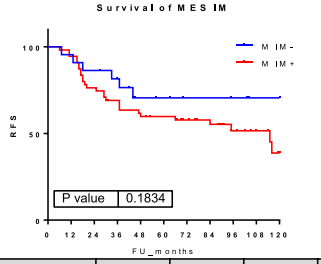
Subjects at risk	0 years	2 years	5 years	8 years	10 years
IM+	84	58	44	23	16
IM-	24	14	11	6	2



Subjects at risk	0 years	2 years	5 years	8 years	10 years
IM+	28	19	16	9	2
IM-	87	54	40	23	18



Subjects at risk	0 years	2 years	5 years	8 years	10 years
IM+	26	18	14	11	5
IM-	30	20	12	9	7

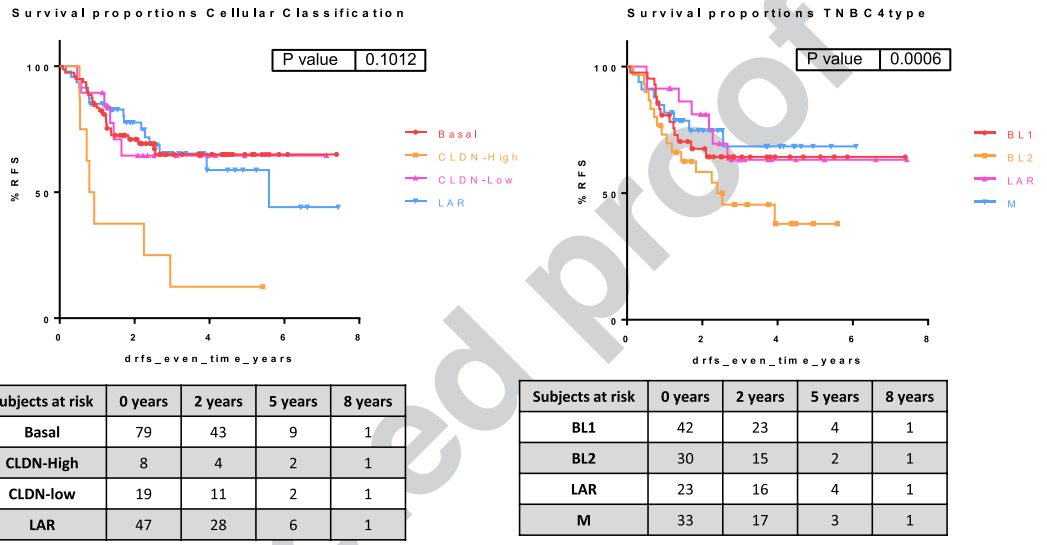


Subjects at risk	0 years	2 years	5 years	8 years	10 years
IM+	55	42	31	15	5
IM-	22	20	8	6	3

Figure 8. Kaplan-Meier survival curves represent the survival rate of immune-positive and immune-negative tumors in the Burstein's subgroups.

Cellular Classification	Number	IM Characteristic	Number	%Intragroup
Basal	79 (52%)	IM−	41	52%
		IM+	38	48%
CLDN-High	8 (5%)	IM−	2	25%
		IM+	6	75%
CLDN-Low	19 (12%)	IM−	12	63%
		IM+	7	37%
LAR	47 (31%)	IM−	25	53%
		IM+	22	47%

Table 7. Shows the cellular classification and the immune characteristic in the neoadjuvant dataset.



Subjects at risk	0 years	2 years	5 years	8 years
Basal	79	43	9	1
CLDN-High	8	4	2	1
CLDN-low	19	11	2	1
LAR	47	28	6	1

Subjects at risk	0 years	2 years	5 years	8 years
BL1	42	23	4	1
BL2	30	15	2	1
LAR	23	16	4	1
M	33	17	3	1

Figure 9. Kaplan–Meier survival curves represent the distant relapse-free survival rate of the cellular and the TNBC4-type subgroups in the GSE25066 series.

The CLDN-high subgroup presented the poorest prognosis among the cellular classification subgroups. Immune-positive tumors had a better prognosis (Fig. 9).

Discussion

TNBC constitutes a heterogeneous disease with various molecular entities. The study of this heterogeneity has thus far not conferred significant advances in the treatment of patients. The application of probabilistic graphical models (PGMs) provides deep insight into high-throughput data[18]. In the present study, we used PGMs to unravel specific molecular information concerning various biological entities, such as the immune status or the developmental point when the breast stem cell turns carcinogenic.

Previous studies used differences in gene expression to define TNBC subtypes[3,6–8,10]. Subtypes emerged from clustering methods such as HCL or non-negative matrix factorization, which group genes around specific functions. On the contrary, we hereby applied an unsupervised analysis, without knowledge of the functions of the genes selected in each step of the process. We ultimately identified the genes involved in 26 different molecular functions, which agreed with the metagenes described by Rody *et al.*[5]. This approach provides two different classifications (immune and cellular), each related to particular genes and functions.

Once the PGM functional structure was established, we defined four subgroups: CLDN-low, CLDN-high, basal-like and LAR, agreeing with the cancer stem cell hypothesis[2,13–15]. These four groups identify the point of the differentiation process where the stem cell becomes carcinogenic: the less differentiated tumors will be CLDN-low, and the most differentiated tumors will be LAR.

Functional node activities confirm that there are differences among cellular subgroups, and some of these differences could have therapeutic utility. For example, the activity of node 19 (PPAR signaling pathway) showed meaningful differences between the CLDN-low subgroup and the other three, suggesting that PPAR-directed therapies might have a different effect on the CLDN-low subgroup. Finally, we observed that cellular subgroups had different clinical features.

On the other hand, the immune layer was described in this study as a compendium of functional nodes, each of which related to a specific immune function. However, when taking all these nodes together as a metanode we were able to establish an immune classification with prognostic value among all the series.

The immune and cellular classifications reflected unrelated biological identities. As shown in Fig. 4, the LAR and CLDN-high subgroups presented different prognoses when split by the immune layer. LAR immune-negative

tumors were associated with a 30% 5-year survival rate compared with 70% in the LAR immune-positive group. The immune-based subtype might also influence the response to immunotherapy. Ongoing trials are evaluating anti-PD1 antibodies in breast cancer, particularly in triple-negative disease[24]. It would be interesting to assess the efficacy of anti-PD1 therapy in subtypes defined by immune layer.

We also compared the cellular classification with other classifications previously described[7,8,10]. LAR is over-represented in every luminal subgroup regardless of the classification, which demonstrates that this is a homogeneous and reproducible group. Similarly, the basal cellular subgroup is overrepresented in basal subgroups across classifications. There is also a high correlation (83%) in the CLDN-low cellular groups, which confirms the existence of a CLDN-low subgroup independent of the expression of ER, PR and HER2, as previously suggested[16].

Our results show that immune features appear across different subtypes. Interestingly, the luminal immune-positive group did much better than the luminal immune-negative group. Regardless of the classification[7,8,10], the immune layer added prognostic information to the luminal subtypes. The immune layer had been previously defined as a separate group in these classifications, but it appears to intersect with other biological features, providing additional prognostic value.

With regard to the cellular classification, our CLDN-low cellular subgroup had an 89% concordance with the basal-like 2 Lehmann's subgroup, which puts BL2 in the stem cell hypothesis context, suggesting that basal-like 2 tumors might be caused by early differentiated carcinogenic stem cells. The CLDN-high subgroup does not appear in other classifications, which suggests that this is an intermediate group between CLDN-low tumors (stem cell not yet expressing CLDN genes) and basal tumors. It might be difficult to draw the line between groups in this continuous, cellular differentiation-based classification, although Burnstein's basal-like immune-active corresponded to the CLDN-high immune-negative in our classification. Regardless of the classification, there was always a luminal subgroup, one or two basal subgroups and some mesenchymal or CLDN subgroup.

Our classification could also provide some predictive information. CLDN-high tumors had a poor response to neoadjuvant chemotherapy. Much effort has been devoted to the prediction of response to chemotherapy in TNBC. Cell-free DNA[25], tumor-infiltrating lymphocytes[26], microRNA signatures[27] and proteomics[28], among others, have recently been proposed as useful methods in this regard. Further research is needed before the cellular classification described in the present paper could be considered in the selection of therapy.

This study has some limitations. The 2010 American Society of Clinical Oncology guidelines established the 1% threshold for the expression of PR and ER[29]; however, our tumor series was assessed before that date, so we cannot ensure that all the TNBC tumors fulfilled this criterion. Another limitation to our study is that the cellular classification is based on a continuum, which makes it difficult to set categories. Finally, these results should be validated in additional cohorts to evaluate the robustness of our cellular and immune classification. However, we believe that our findings serve as an important hypothesis in generating findings that can be explored in future studies.

Conclusion

In conclusion, the use of probabilistic graphical models in TNBC suggests that there are at least two independent biological layers, cellular and immune. We propose a new way to characterize TNBC taking these two dimensions into account, and leading to the result that the luminal immune-positive subgroup had a better prognosis than the luminal immune-negative.

Availability of Data and Material

The datasets analyzed during the current study, GSE31519 [https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE31519], and GSE25066 [https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE25066], are available in the GEO Datasets repository.

References

1. Network, C. G. A. Comprehensive molecular portraits of human breast tumours. *Nature* **490**, 61–70 (2012).
2. Stingl, J. & Caldas, C. Molecular heterogeneity of breast carcinomas and the cancer stem cell hypothesis. *Nat Rev Cancer* **7**, 791–799 (2007).
3. Perou, C. M. *et al*. Molecular portraits of human breast tumours. *Nature* **406**, 747–752 (2000).
4. Yersal, O. & Barutca, S. Biological subtypes of breast cancer: Prognostic and therapeutic implications. *World J Clin Oncol* **5**, 412–424 (2014).
5. Rody, A. *et al*. A clinically relevant gene signature in triple negative and basal-like breast cancer. *Breast Cancer Res* **13**, R97 (2011).
6. Lehmann, B. D. *et al*. Identification of human triple-negative breast cancer subtypes and preclinical models for selection of targeted therapies. *J Clin Invest* **121**, 2750–2767 (2011).
7. Lehmann, B. D. *et al*. Refinement of Triple-Negative Breast Cancer Molecular Subtypes: Implications for Neoadjuvant Chemotherapy Selection. *PLoS One* **11**, e0157368 (2016).
8. Burstein, M. D. *et al*. Comprehensive genomic analysis identifies novel subtypes and targets of triple-negative breast cancer. *Clin Cancer Res* **21**, 1688–1698 (2015).
9. Sabatier, R. *et al*. Kinome expression profiling and prognosis of basal breast cancers. *Mol Cancer* **10**, 86 (2011).
10. Prat, A. & Perou, C. M. Deconstructing the molecular portraits of breast cancer. *Mol Oncol* **5**, 5–23 (2011).
11. Jézéquel, P. *et al*. Gene-expression signature functional annotation of breast cancer tumours in function of age. *BMC Med Genomics* **8**, 80 (2015).
12. Milioli, H. H., Tishchenko, I., Riveros, C., Berretta, R. & Moscato, P. Basal-like breast cancer: molecular profiles, clinical features and survival outcomes. *BMC Med Genomics* **10**, 19 (2017).
13. Shipitsin, M. & Polyak, K. The cancer stem cell hypothesis: in search of definitions, markers, and relevance. *Lab Invest* **88**, 459–463 (2008).
14. Sims, A. H., Howell, A., Howell, S. J. & Clarke, R. B. Origins of breast cancer subtypes and therapeutic implications. *Nat Clin Pract Oncol* **4**, 516–525 (2007).
15. Allegra, A. *et al*. The cancer stem cell hypothesis: a guide to potential molecular targets. *Cancer Invest* **32**, 470–495 (2014).

16. Prat, A. *et al*. Phenotypic and molecular characterization of the claudin-low intrinsic subtype of breast cancer. *Breast Cancer Res* **12**, R68 (2010).
17. Ritchie, M. E. *et al*. limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Res* **43**, e47 (2015).
18. Gámez-Pozo, A. *et al*. Combined Label-Free Quantitative Proteomics and microRNA Expression Analysis of Breast Cancer Unravel Molecular Differences with Clinical Implications. *Cancer Res* **75**, 2243–2253 (2015).
19. Huang, D. W., Sherman, B. T. & Lempicki, R. A. Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nat Protoc* **4**, 44–57 (2009).
20. de Velasco, G. *et al*. Urothelial cancer proteomics provides both prognostic and functional information. *Sci Rep* **7**, 15819 (2017).
21. Monti, S., Tamayo, P., Mesirov, J. & Golub, T. Consensus Clustering: A Resampling-Based Method for Class Discovery and Visualization of Gene Expression Microarray Data. *Machine learning* **52**, 91–118 (2003).
22. Witten, D. M. & Tibshirani, R. A framework for feature selection in clustering. *J Am Stat Assoc* **105**, 713–726 (2010).
23. Saeed, A. I. *et al*. TM4: a free, open-source system for microarray data management and analysis. *Biotechniques* **34**, 374–378 (2003).
24. Katz, H. & Alsharedi, M. Immunotherapy in triple-negative breast cancer. *Med Oncol* **35**, 13 (2017).
25. Sung, J. S. *et al*. Detection of somatic variants and. *Oncotarget* **8**, 106901–106912 (2017).
26. Wein, L. *et al*. Clinical Validity and Utility of Tumor-Infiltrating Lymphocytes in Routine Clinical Practice for Breast Cancer Patients: Current and Future Directions. *Front Oncol* **7**, 156 (2017).
27. García-Vazquez, R. *et al*. A microRNA signature associated with pathological complete response to novel neoadjuvant therapy regimen in triple-negative breast cancer. *Tumour Biol* **39**, 1010428317702899 (2017).
28. Gámez-Pozo, A. *et al*. Prediction of adjuvant chemotherapy response in triple negative breast cancer with discovery and targeted proteomics. *PLoS One* **12**, e0178296 (2017).
29. Hammond, M. E., Hayes, D. F., Wolff, A. C., Mangu, P. B. & Temin, S. American society of clinical oncology/college of american pathologists guideline recommendations for immunohistochemical testing of estrogen and progesterone receptors in breast cancer. *J Oncol Pract* **6**, 195–197 (2010).

Acknowledgements

Q7
This study was funded by Instituto de Salud Carlos III, Spanish Economy and Competitiveness Ministry, Spain and co-funded by the FEDER program, "Una forma de hacer Europa" (PI15/01310). LT-F is supported by the Spanish Economy and Competitiveness Ministry (DI-15-07614). GP-V is supported by Conserjería de Educación, Juventud y Deporte of Comunidad de Madrid (IND2017/BMD7783).

Author Contributions

All the authors have directly participated in the preparation of this manuscript and have read and approved the final version submitted. J.M.A., H.N. and P.M. contributed the probabilistic graphical models. M.D.-A. and G.P.-V. contributed the statistical analyses. G.P.-V., L.T.-F., A.Z.-M., M.F.-G. and R.L.-V. performed the probabilistic graphical model interpretation and the gene ontology analyses. G.P.-V. drafted the manuscript. G.P.-V., A.G.-P., J.A.F.V., J.F., P.Z. and E.E. conceived of the study and participated in its design and interpretation. A.G.-P., J.A.F.V., E.E. and L.T.-F. supported the manuscript drafting. A.G.-P. and J.A.F.V. coordinated the study.

Additional Information

Supplementary information accompanies this paper at https://doi.org/10.1038/s41598-018-38364-y.

Competing Interests: A.F.V., A.G.-P. and E.E. are shareholders of Biomedica Molecular Medicine S.L. L.T.-F. and G.P.-V. are employees of Biomedica Molecular Medicine S.L. The other authors declare no competing interests.

Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Melanoma proteomics unravels major differences related to mutational status

Lucía Trilla-Fuertes[1#], Angelo Gámez-Pozo[1,2#], Guillermo Prado-Vázquez[1,2], Andrea Zapater-Moros[1,2], Mariana Díaz-Almirón[3], Claudia Fortes[4], María Ferrer-Gómez[2], Rocío López-Vacas[2], Verónica Parra Blanco[5], Iván Márquez-Rodas[6,9], Ainara Soria[7], Juan Ángel Fresno Vara[2,9,$] and Enrique Espinosa[8,9,$]

[1]Biomedica Molecular Medicine SL, Madrid, Spain

[2]Molecular Oncology & Pathology Lab, Institute of Medical and Molecular Genetics-INGEMM, Hospital Universitario La Paz-IdiPAZ, Madrid, Spain

[3]Biostatistics Unit, Hospital Universitario La Paz-IdiPAZ, Madrid, Spain

[4]Functional Genomics Center Zurich, University of Zurich/ETH Zurich, Zurich, Switzerland

[5]Servicio de Anatomía Patológica, Hospital Universitario Gregorio Marañón, Madrid, Spain

[6]Servicio de Oncología Médica, Hospital Universitario Gregorio Marañón, Madrid, Spain

[7]Servicio de Oncología Médica, Hospital Universitario Ramón y Cajal, Madrid, Spain

[8]Servicio de Oncología Médica, Hospital Universitario La Paz-IdiPAZ, Madrid, Spain

[9] Biomedical Research Networking Center on Oncology-CIBERONC, ISCIII, Madrid, Spain

[#]These authors contributed equally to this work

[$] Co-Corresponding Authors

Co-Corresponding Authors: Juan Ángel Fresno Vara. juanangel.fresno@salud.madrid.org.

Enrique Espinosa Arranz . eespinosa00@hotmail.com

Abstract

Melanoma is the most lethal cutaneous cancer. New drugs have recently appeared; however, not all patients obtain a benefit of these new drugs. For this reason, it is still necessary to characterize melanoma at molecular level. The aim of this study was to explore the molecular differences between melanoma tumor subtypes, based on BRAF and NRAS mutational status. Fourteen formalin-fixed, paraffin-embedded melanoma samples were analyzed using a high-throughput proteomics approach, coupled with probabilistic graphical models and Flux Balance Analysis, to characterize these differences. Proteomics analyses showed differences in expression of proteins related with fatty acid metabolism, melanogenesis and extracellular space between BRAF mutated and BRAF non-mutated melanoma tumors. Additionally, probabilistic graphical models showed differences between melanoma subgroups at biological processes such as melanogenesis or metabolism. On the other hand, Flux Balance Analysis predicts a higher tumor growth rate in BRAF mutated melanoma samples. In conclusion, differential biological processes between melanomas showing a specific mutational status can be detected using combined proteomics and computational approaches.

Introduction

Melanoma is the most lethal cutaneous cancer, with over 11,000-15,000 estimated deaths in the United States and Europe every year [1,2]. Better understanding of the molecular biology of this tumor has allowed the development of new effective drugs for the treatment of advanced disease, both in the fields of targeted therapies and immunotherapy [3]. However, as not all patients obtain a benefit from new drugs, further insight into the biology of melanoma is needed.

Gene signatures, genomic hybridization, whole-exome genome sequencing, microRNA analysis and other techniques have widely addressed the genomic landscape of melanoma, contributing to significant advances [4,5]. Given the heterogeneity of melanoma and the complex interaction of this tumor with the immune system, the need for combination of biomarkers assays has been recently proposed to properly analyze the disease [6].

Proteins determine cell phenotype, so proteomics analyses offer the possibility to measure the biologic outcome of cancer-related genomic abnormalities [7]. Mass spectrometry has become the method of choice to assess complex protein samples, and recent technological advances allow the identification of thousands of proteins from tissue amounts compatible with clinical routine. Therefore, proteomics may become a new source of molecular cancer markers offering complementary information to that provided by standard pathology and genomics. We recently demonstrated the feasibility of high-throughput label-free quantitative proteomics to analyze breast cancer from paraffin-embedded samples [8]. In the present study we sought to determine whether high-throughput proteomics combined with computational approaches, such as probabilistic graphical models and Flux Balance Analysis, are useful tools to explore functional differences between groups of melanoma tumors.

Results

Patients and samples

Primary melanoma samples coming from 14 patients with advanced disease were included. Samples were split into three groups according to mutational status: BRAF-mutant (n=3), NRAS-mutant (n= 5) or double negative (n= 6). BRAF and NRAS mutations had been previously determined in local laboratories with standard polymerase chain reaction-based tests.

Mass-spectrometry analysis

FFPE melanoma tumor samples were analysed by mass-spectrometry. 4,006 protein groups were identified, of which 1,606 present at least two unique peptides and detectable expression in at least 75% of the samples. Label-free quantification values from these 1,606 proteins were used for subsequent analyses.

Differential protein expression patterns between subtypes

A Significance Analysis of Microarrays (SAM) was done to find differences among samples at the protein level. Seventeen proteins were found differentially expressed between BRAF mutated and BRAF wild type tumors, all of them underexpressed in BRAF-mutated tumors (Fig 1). These proteins are mainly related with fatty acid metabolism.

In addition, delta values between BRAF-mutated and BRAF-wild type, and NRAS-mutated and NRAS-wild type tumors were calculated. Delta values higher than 1.5 or lower than -1.5 were used to perform gene ontology analyses as well. Proteins related with keratinization, epidermis development and cytoskeleton were underexpressed, whereas proteins involved in melanogenesis and extracellular space were overexpressed in BRAF-mutant as compared with BRAF-wild type samples. SAM and delta analyses did not find significant differences between NRAS-mutant and NRAS-wild type tumors.

Probabilistic graphical model and node activity measurements

A probabilistic graphical model was built using proteomics data without other *a priori* information. The resulting network was processed to build a functional structure, as described in previous works [14-16]. The resulting network was divided into thirteen branches, and gene ontology analyses were performed to establish functional structure. Finally, twelve principal functions were assigned to different branches and there was a branch to which no main function could be assigned (Fig 2).

Node activity measurements were calculated for each node using proteins related with the main assigned function and a comparison between BRAF-mutant, NRAS-mutant and double-negative groups was performed. Although the limited number of samples did not allow seeing significant differences, some trends in functional activities were found. For instance, NRAS-mutant had a lower melanosome node activity than BRAF-mutant or double negative tumors. On the other hand, BRAF-mutant tumors had a higher metabolism node activity than NRAS-mutant or double negative (Fig 3).

Flux Balance Analysis

Flux Balance Analysis is a computational approach to assess biochemical networks through the calculation of the flow of metabolites through this network. FBA can be used to calculate the growth rate of an organism or a tumor or the rate of production of a given metabolite. Our model predicted that BRAF mutated tumors have a higher tumor growth rate than the two other subtypes (Fig 4).

Discussion

In this study, proteomics coupled with probabilistic graphical models and flux balance analysis were used to characterize differences between melanoma biomarker subgroups in melanoma samples.

Mass-spectrometry workflow allowed the detection of 1,606 proteins with two unique peptides and detectable expression in at least 75% of the samples. Differences in fatty acid metabolism, cytoskeleton or keratinization were observed between BRAF–mutant and BRAF-wild type tumors. Also, differences in functions such as melanogenesis or metabolism were shown between subgroups

SAM and gene ontology analysis found 17 proteins differentially expressed between BRAF-mutant and the two other subgroups (NRAS-mutant and double-negative), all of them were underexpressed in BRAF-mutated tumors. These proteins are mainly involved in fatty acid metabolism: acyl-Co A dehydrogenases P11310 (acyl-CoA dehydrogenase medium chain, ACADM), P42765 (acetyl-CoA acyltransferase 2, ACAA2) and P49748 (acyl-CoA dehydrogenase very long chain, ACADVL). On the other hand, some of the proteins underexpressed in BRAF-mutated tumors have antiproliferative functions. For instance, Q9Y3Z3 (histidine/aspartate (HD)- domain containing protein 1, SAMHD1) is implicated in regulation of DNA replication and damage repair and it is proposed to have antiproliferative and tumor suppressive functions in many cancers [20]. Q96CX2 (potassium channel tetradimerization domain containing 12, KCTD12) inhibits proliferation in uveal melanoma cells [21]. O14745 (SLC9A3R1) is involved in suppressing breast cancer cells proliferation [22]. Other proteins of those 17 were previously related with melanoma or melanogenesis processes. For example, P31040 (succinate dehydrogenase complex flavoprotein subunit A,SDHA), which encodes a major catalytic subunit of succinate-ubiquinone reductase, a complex of the mitochondria chain, it was previously related with melanogenesis process [23]. Another protein differentially expressed is

P00488 (coagulation factor XIII, F13A1) which it was previously associated with chemotherapy response in melanoma tumors [24]. P08133 (annexin A6, ANXA6) acts as a tumor suppressor in skin cancer and it is involved in in the conversion of melanocytes to malignant melanomas [25]. Lastly, it was previously described that metastatic melanoma tumors have a decreased expression of signal transducer and activator of transcription P42224 (signal transducer and activator of transcription 1, STAT1) and it could be one of the mechanism by which melanoma can evade immune detection [26]. Finally, P04899 (G protein subunit alpha i2, GNAI2) contributes to melanoma cell growth [27]. O60749 (Sorting nexin 2, SNX2) is involved in membrane trafficking of growth factor receptors including epidermal growth factor receptor and c-Met [28]. P05107 (integrin subunit beta 2, ITGB2) participates in cell adhesion as well as cell-surface mediated signalling and it is correlated with survival in other cancers such as renal or colorectal tumors [29,30]. As far we know, P14317 (hematopoietic cell-specific Lyn substrate, HCLS1), P63027 (vesicle-associated membrane protein 2, VAMP2), P16402 (histone cluster 1H1 family member d, HIST1H1D) and Q9H0W9 (chromosome 11 open reading frame 54, C11orf54) were not previously related with melanoma or other cancers.

Differential analyses did not show differences between NRAS-mutant and NRAS-wild type tumors, which are attributable to the small sample size. The present study was limited in this regard because it was designed just as a proof of principle that high-throughput proteomics can be used to study clinical samples of melanoma. Future studies with larger sample size will be needed to establish significant differences among subtypes. Interestingly, it seems that delta analyses and SAM provide complementary information about different protein expression patterns, because differential proteins provided by these two analyses were different and they were also related to different biological processes.

On the other hand, a probabilistic graphical model was used to generate a network based in protein expression data. It is remarkable that, despite the low number of samples, the

probabilistic graphical model clearly showed a functional structure. This type of analysis previously demonstrated its utility to characterize other tumor types such as bladder carcinoma or breast cancer and may complement the information provided by genomics [15,16]. On the other hand, the high growth rate in BRAF-mutant tumors predicted by FBA agrees with previous knowledge as BRAF-mutated tumors are more proliferative [31].

Our study demonstrates that proteomics and computational methods can be applied to the study of formalin-fixed, paraffin-embedded melanoma samples, suggesting that melanoma subgroups, defined by mutational status, have major molecular differences. Despite the reduced number of samples analyzed, the probabilistic graphical model showed a functional structure and allowed characterizing differences at biological processes regarding melanoma mutational status. Additionally, flux balance analysis was capable to predict differences at tumor growth rate between these groups. In conclusion, proteomics and computational approaches demonstrated their usefulness in the molecular characterization of melanoma and suggested some proteins and biological processes that could be used as therapeutic targets.

Material and Methods

Samples

Fourteen melanoma cancer patients were included in the study. FFPE samples were retrieved from Biobanks in IdiPAZ, Hospital Universitario Gregorio Marañón and Hospital Universitario Ramón y Cajal, all integrated in the Spanish Hospital Biobank Network (RetBioH; http://www.redbiobancos.es/). Patients provided informed consent. All experiments were performed in accordance with relevant guidelines and regulations. The histopathological features of each sample were reviewed by an experienced pathologist to confirm diagnosis and tumor content. Eligible samples had to include at least 50% of tumor cells. Approval from the Ethical Committees of Hospital Universitario La Paz was obtained for the conduct of the study.

Mass-spectrometry analysis protein identification and label-free quantification

Proteins were extracted from FFPE samples as previously described [9]. Peptides were desalted using self-packed C18 stage tips, dried and resolubilized with 15µl of 3% acetonitrile, 0.1% formic acid. Mass spectrometry analysis was performed on a QExactive mass spectrometer coupled to a nano EasyLC 1000 (Thermo Fisher Scientific). Solvent composition at the two channels was 0.1% formic acid for channel A and 0.1% formic acid, 99.9% acetonitrile for channel B. For each sample 3µL of peptides were loaded on a commercial PepMapTM RSLC C18 Snail Column (75 µm × 500 mm, Thermo Fisher Scientific) and eluted at a flow rate of 300 nL/min by a gradient from 2 to 30% B in 85 min, 47% B in 4 min and 98% B in 4 min. Samples were acquired in a randomized order. The mass spectrometer was operated in data-dependent mode (DDA), acquiring a full-scan MS spectra (300−1700 m/z) at a resolution of 70000 at 200 m/z after accumulation to a target value of 3000000, followed by HCD (higher-energy collision dissociation) fragmentation on the twelve most intense signals per cycle. HCD spectra were acquired at a resolution of 35000 using normalized collision energy of 25 and a

maximum injection time of 120 ms. The automatic gain control (AGC) was set to 50000 ions. Charge state screening was enabled and singly and unassigned charge states were rejected. Only precursors with intensity above 8300 were selected for MS/MS (2% underfill ratio). Precursor masses previously selected for MS/MS measurement were excluded from further selection for 30 s, and the exclusion window was set at 10 ppm. The samples were acquired using internal lock mass calibration on m/z 371.1010 and 445.1200.

The acquired raw MS data were processed by MaxQuant (version 1.5.2.8), followed by protein identification using the integrated Andromeda search engine. Spectra were searched against a forward Swiss Prot-human database, concatenated to a reversed decoyed fasta database and common protein contaminants (NCBI taxonomy ID9606, release date 2014-05-06). Carbamidomethylation of cysteine was set as fixed modification, while methionine oxidation and N-terminal protein acetylation were set as variable. Enzyme specificity was set to trypsin/P allowing a minimal peptide length of 7 amino acids and a maximum of two missed-cleavages. Precursor and fragment tolerance was set to 10 ppm and 20 ppm, respectively for the initial search. The maximum false discovery rate (FDR) was set to 0.01 for peptides and 0.05 for proteins. Label free quantification was enabled and a 2 minutes window for match between runs was applied. The re-quantify option was selected. For protein abundance the intensity was used, corresponding to the sum of the precursor intensities of all identified peptides for the respective protein group (Sup Table 1).

Following MS workflow, identified protein groups were filtered by the presence of at least two unique peptides and detectable expression in at least 75% of the samples. Label-free quantification values from these proteins were used for subsequent analyses. Additionally, batch effects were removed using *limma* package [10] and R v 3.2.5 [11].

Protein differential expression analyses

Significance Analysis of Microarrays (SAM) was performed using MeV to find significant differences in protein expression among samples [12]. Protein expression patterns were also compared calculating delta values for each biomarker status against the rest of the tumor samples. Proteins showing a change in expression value higher than 1.5 or lower than -1.5 were selected.

Probabilistic graphical model and activity measurements

R v 3.2.5 [11] and *grapHD* package [13] were used to build a probabilistic graphical model using correlation coefficient as associative method as previously described [14-16]. Protein expression data was used to build the network with any *a priori* information. The network was split into several branches and Gene Ontology analysis was used to assign a major function to each branch, dividing the network into functional nodes. Activity measurements were then calculated by the mean expression of all the proteins related to the assigned node function.

Flux Balance Analysis

Flux Balance Analysis (FBA) was performed using COBRA Toolbox [17] and whole metabolism human reconstruction Recon 2 [18] both available for MATLAB. As an objective function, biomass reaction supplied by the Recon 2 was used as representative of tumor growth rate. Proteomics expression data was incorporated into the model as described in previous works [15]. Briefly, Gene-Protein-Reaction (GPR) rules were estimated using the sum for "ORs" expressions and minimum for "ANDs" expressions. Then, E-flux algorithm [19] was used to normalize the GPR values dividing by the maximum value in each tumor and incorporate protein expression data into the model.

Statistical analyses

GraphPad Prism v6 was used for statistical analyses, whereas Cytoscape was used for network analysis. Gene Ontology Analyses were performed in DAVID webtool selecting "Homo sapiens" background and GOTERM-FAT, Biocarta and KEGG databases.

Acknowledgements

We want to particularly acknowledge the patients in this study for their participation and to IdiPAZ, as well as participating Biobanks. LT-F is supported by the Spanish Economy and Competitiveness Ministry (DI-15-07614). GP-V is supported by Conserjería de Educación, Juventud y Deporte of Comunidad de Madrid (IND2017/BMD7783). This work was supported by Instituto de Salud Carlos III, Spanish Economy and Competitiveness Ministry, Spain and co-funded by the FEDER program, "Una forma de hacer Europa" (PI15/01310). The funders had no role in the study design, data collection and analysis, decision to publish or preparation of the manuscript.

Author Contributions:

All the authors have directly participated in the preparation of this manuscript and have approved the final version submitted and declare no ethical conflicts of interest. R.L.-V. contributed the protein extraction. C.F contributed the mass spectrometry data. G.P-V, A.Z-M. and M.F-G. contributed the probabilistic graphical models. M.D.-A. contributed the GPR rule method. V.P-B., I.M-R, A.S. and E.E. contributed the clinical data and the analyses related. A.G.-P., G.P.-V., A.Z.-M., and L.T-F. contributed in the design of the study and the statistical and gene ontology analyses. L.T.-F. drafted the manuscript. L.T.-F. contributed the FBA analyses. J.A.F.V., A.G-P. and E.E. conceived of the study, and participated in its design and interpretation. J.A.F.V. and E.E. coordinated the study. All authors read and approved the final manuscript..

Conflict of interest: JAFV, EE and AG-P are shareholders in Biomedica Molecular Medicine SL. LT-F and GP-V are employees of Biomedica Molecular Medicine SL. The other authors declare no competing interests.

References

1 Ferlay, J. *et al.* Cancer incidence and mortality patterns in Europe: estimates for 40 countries in 2012. *Eur J Cancer* **49**, 1374-1403, doi:10.1016/j.ejca.2012.12.027 (2013).
2 Ferlay, J. *et al.* Cancer incidence and mortality worldwide: sources, methods and major patterns in GLOBOCAN 2012. *Int J Cancer* **136**, E359-386, doi:10.1002/ijc.29210 (2015).
3 Dummer, R., Keilholz, U. & Committee, E. G. appendix 2: Cutaneous melanoma (2): eUpdate published online September 2016 (http://www.esmo.org/Guidelines/Melanoma). *Ann Oncol* **27**, v136-v137, doi:10.1093/annonc/mdw432 (2016).
4 Lin, W. M. & Fisher, D. E. Signaling and Immune Regulation in Melanoma Development and Responses to Therapy. *Annu Rev Pathol* **12**, 75-102, doi:10.1146/annurev-pathol-052016-100208 (2017).
5 Bauer, J. The Molecular Revolution in Cutaneous Biology: Era of Cytogenetics and Copy Number Analysis. *J Invest Dermatol* **137**, e57-e59, doi:10.1016/j.jid.2016.11.043 (2017).
6 Blank, C. U., Haanen, J. B., Ribas, A. & Schumacher, T. N. CANCER IMMUNOLOGY. The "cancer immunogram". *Science* **352**, 658-660, doi:10.1126/science.aaf2834 (2016).
7 Ellis, M. J. *et al.* Connecting genomic alterations to cancer biology with proteomics: the NCI Clinical Proteomic Tumor Analysis Consortium. *Cancer Discov* **3**, 1108-1112, doi:10.1158/2159-8290.CD-13-0219 (2013).
8 Gámez-Pozo, A. *et al.* Prediction of adjuvant chemotherapy response in triple negative breast cancer with discovery and targeted proteomics. *PLoS One* **12**, e0178296, doi:10.1371/journal.pone.0178296 (2017).
9 Gámez-Pozo, A. *et al.* Shotgun proteomics of archival triple-negative breast cancer samples. *Proteomics Clin Appl* **7**, 283-291, doi:10.1002/prca.201200048 (2013).
10 Ritchie, M. *et al.* Vol. 43 e47 (Nucleic Acid Research, 2015).
11 (Vienna,Austria. R Foundation for Stattistical Computing, 2013).
12 Saeed, A. I. *et al.* TM4: a free, open-source system for microarray data management and analysis. *Biotechniques* **34**, 374-378 (2003).
13 Abreu, G., Edwards, D. & Labouriau, R. Vol. 37 1-18 (Journal of Statistical Software, 2010).
14 Gámez-Pozo, A. *et al.* Vol. 75 2243-2253 (Cancer Res, 2015).
15 Gámez-Pozo, A. *et al.* Functional proteomics outlines the complexity of breast cancer molecular subtypes. *Scientific Reports* **7**, 10100, doi:10.1038/s41598-017-10493-w (2017).
16 de Velasco, G. *et al.* Urothelial cancer proteomics provides both prognostic and functional information. *Sci Rep* **7**, 15819, doi:10.1038/s41598-017-15920-6 (2017).
17 Schellenberger, J. *et al.* Vol. 6 1290-1307 (Nature Protocols, 2011).
18 Thiele, I. *et al.* A community-driven global reconstruction of human metabolism. *Nat Biotechnol* **31**, 419-425, doi:10.1038/nbt.2488 (2013).
19 Colijn, C. *et al.* Vol. 5 (PLOS Comput Bio, 2009).
20 Kohnken, R., Kodigepalli, K. M. & Wu, L. Regulation of deoxynucleotide metabolism in cancer: novel mechanisms and therapeutic implications. *Mol Cancer* **14**, 176, doi:10.1186/s12943-015-0446-6 (2015).
21 Luo, L. *et al.* Lentiviral-mediated overexpression of KCTD12 inhibits the proliferation of human uveal melanoma OCM-1 cells. *Oncol Rep* **37**, 871-878, doi:10.3892/or.2016.5325 (2017).
22 Liu, H. *et al.* SLC9A3R1 stimulates autophagy via BECN1 stabilization in breast cancer cells. *Autophagy* **11**, 2323-2334, doi:10.1080/15548627.2015.1074372 (2015).

23 Boulton, S. J. & Birch-Machin, M. A. Impact of hyperpigmentation on superoxide flux and melanoma cell metabolism at mitochondrial complex II. *FASEB J* **29**, 346-353, doi:10.1096/fj.14-261982 (2015).

24 Azimi, A. *et al.* Proteomics analysis of melanoma metastases: association between S100A13 expression and chemotherapy resistance. *Br J Cancer* **110**, 2489-2495, doi:10.1038/bjc.2014.169 (2014).

25 Qi, H. *et al.* Role of annexin A6 in cancer. *Oncol Lett* **10**, 1947-1952, doi:10.3892/ol.2015.3498 (2015).

26 Osborn, J. L. & Greer, S. F. Metastatic melanoma cells evade immune detection by silencing STAT1. *Int J Mol Sci* **16**, 4343-4361, doi:10.3390/ijms16024343 (2015).

27 Hermouet, S., Aznavoorian, S. & Spiegel, A. M. In vitro and in vivo growth inhibition of murine melanoma K-1735 cell by a dominant negative mutant alpha subunit of the Gi2 protein. *Cell Signal* **8**, 159-166 (1996).

28 Ogi, S. *et al.* Sorting nexin 2-mediated membrane trafficking of c-Met contributes to sensitivity of molecular-targeted drugs. *Cancer Sci* **104**, 573-583, doi:10.1111/cas.12117 (2013).

29 Boguslawska, J. *et al.* Expression of Genes Involved in Cellular Adhesion and Extracellular Matrix Remodeling Correlates with Poor Survival of Patients with Renal Cancer. *J Urol* **195**, 1892-1902, doi:10.1016/j.juro.2015.11.050 (2016).

30 Cavalieri, D. *et al.* Analysis of gene expression profiles reveals novel correlations with the clinical course of colorectal cancer. *Oncol Res* **16**, 535-548 (2007).

31 Wellbrock, C. *et al.* Oncogenic BRAF regulates melanoma proliferation through the lineage specific factor MITF. *PLoS One* **3**, e2734, doi:10.1371/journal.pone.0002734 (2008).

Figure legends

Figure 1:

Differential proteins obtained by Significance Analysis of Microarrays between BRAF positive and negative tumors. 17 proteins were found differentially expressed between BRAF positive and BRAF negative tumors (green= underexpressed, red = overexpressed).

Figure 2:

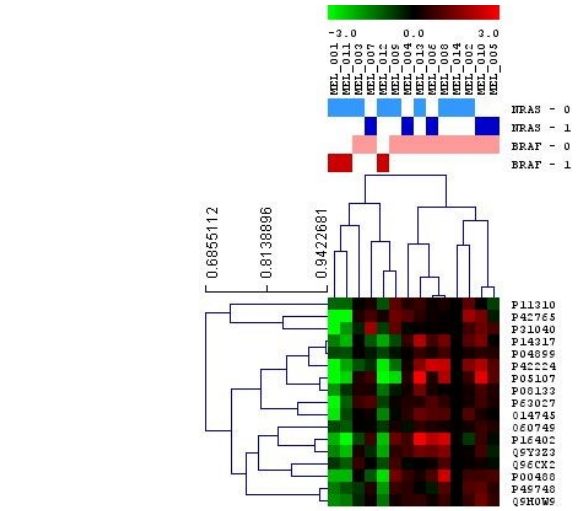
Probabilistic graphical model built using protein expression data from melanoma tumors which showed a functional structure.

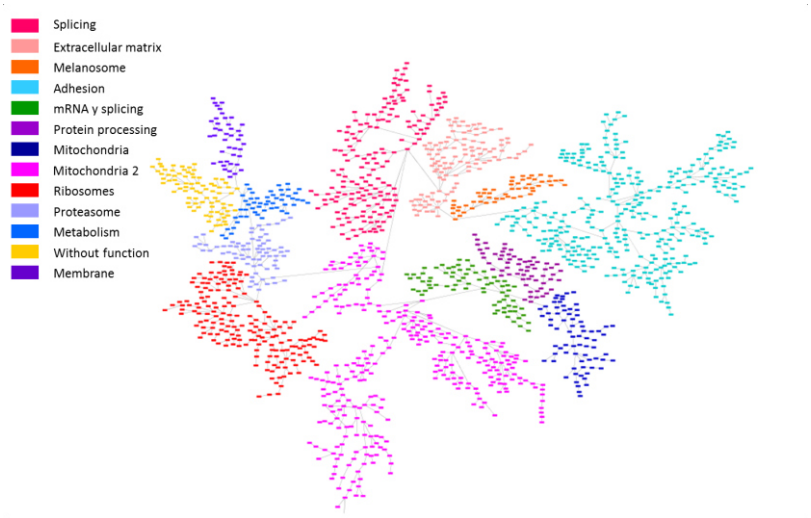
Figure 3:

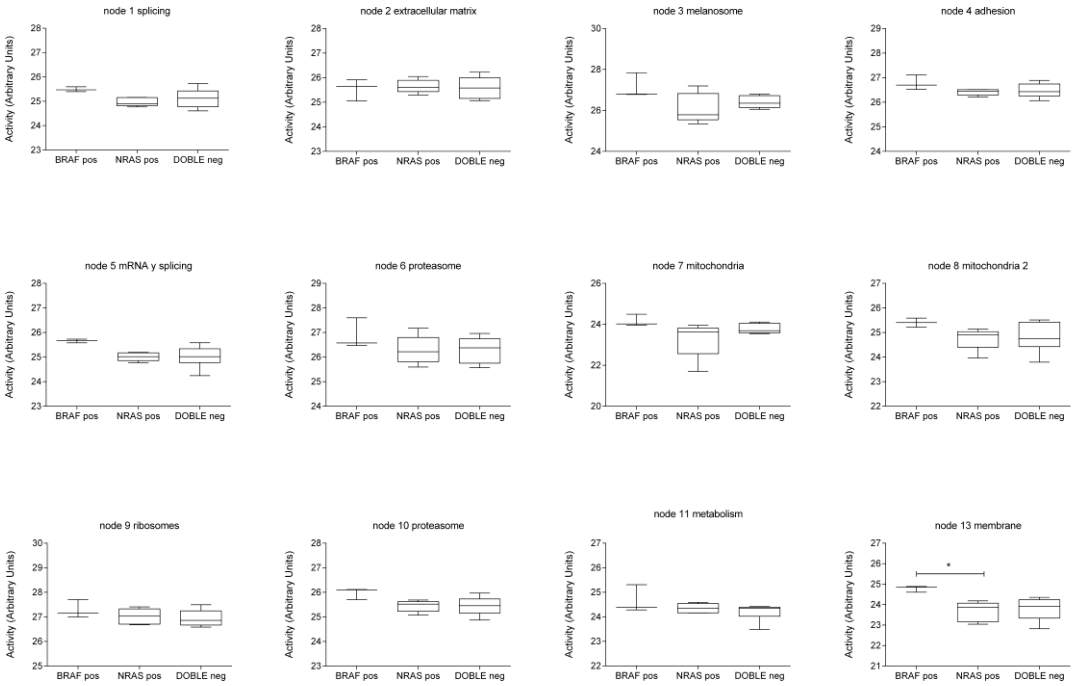
Activity measurements calculated for each network functional node according to biomarkers features.

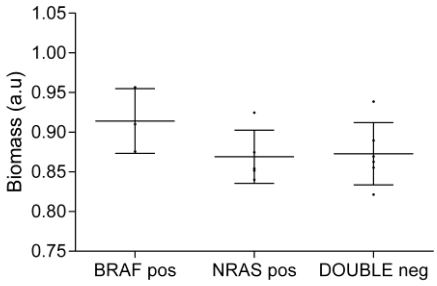
Figure 4:

FBA predicted growth rates. FBA predicted a higher growth rate for BRAF mutated tumors.









Novel Molecular Classification of Muscle-Invasive Bladder Cancer Opens New Treatment Opportunities

Lucía Trilla-Fuertes[1#] BSc, Angelo Gámez-Pozo[1,2#] PhD, Guillermo Prado-Vázquez[1] BSc, Andrea Zapater-Moros[1,2] BSc, Mariana Díaz-Almirón[3] PhD, Jorge M Arevalillo[4] Prof., María Ferrer-Gómez[2] BSc, Hilario Navarro[4] Prof., Paloma Maín[5] Prof., Enrique Espinosa[6,7] MD, Álvaro Pinto[6,7] MD, and Juan Ángel Fresno Vara[2,7,$] PhD

[1]Biomedica Molecular Medicine SL, Madrid, Spain

[2]Molecular Oncology & Pathology Lab, Institute of Medical and Molecular Genetics-INGEMM, Hospital Universitario La Paz-IdiPAZ, Madrid, Spain

[3]Biostatistics Unit, Hospital Universitario La Paz-IdiPAZ, Madrid, Spain

[4]Department of Statistics, Operational Research and Numerical Analysis, Universidad Nacional de Educación a Distancia (UNED).

[5]Department of Statistics and Operations Research, Faculty of Mathematics, Complutense University of Madrid, Madrid, Spain.

[6]Servicio de Oncología Médica, Hospital Universitario La Paz-IdiPAZ, Madrid, Spain

[7]Biomedical Research Networking Center on Oncology-CIBERONC, ISCIII, Madrid, Spain

[$] Corresponding Author

Juan Ángel Fresno Vara, Paseo de la Castellana 261 Madrid, 28046, Spain, Phone: +34912071010 ext255, juanangel.fresno@salud.madrid.org

Word count of the abstract: 241 words

Word count of text: 2820 words

Keywords: computational analyses, immune status, molecular subtypes, muscle-invasive bladder cancer, personalized medicine.

Abstract

Background: Muscle-invasive bladder tumors are associated with high risk of relapse and metastasis even after neoadjuvant chemotherapy and radical cystectomy. Therefore, further therapeutic options are needed and molecular characterization of the disease may help to identify new targets.

Objective: The aim of this work is to characterize muscle-invasive bladder tumors at molecular levels using computational analyses.

Design, Settings and Participants: The TCGA cohort of muscle-invasive bladder cancer patients was used to describe these tumors.

Outcome Measurements and Statistical Analysis: Probabilistic graphical models, layer analyses based on sparse k-means coupled with Consensus Cluster, and Flux Balance Analysis were applied to characterize muscle-invasive bladder tumors at functional level.

Results: Luminal and Basal groups were identified, and an immune molecular layer with independent value was also described. Luminal tumors had decreased activity in the nodes of epidermis development and extracellular matrix, and increased activity in the node of steroid metabolism leading to a higher expression of androgen receptor. This fact points to androgen receptor as a therapeutic target in this group. Basal tumors were highly proliferative according to Flux Balance Analysis, which make these tumors good candidates for neoadjuvant chemotherapy. Immune-high group had higher expression of immune biomarkers, suggesting that this group may benefit from immune therapy.

Conclusions: Our approach, based on layer analyses, established a Luminal group candidate for androgen receptor inhibitor therapy, a proliferative Basal group which seems to be a good candidate for chemotherapy, and an immune-high group candidate for immunotherapy.

Patient Summary: Muscle-invasive bladder cancer has a poor prognosis in spite of appropriate therapy. Therefore, it is still necessary to characterize these tumors to propose new therapeutic targets. In this work we used computational analyses to characterize these tumors and propose treatments.

Introduction

Bladder cancer is estimated to account for 81,190 new cases and 17,240 deaths in the United States in 2018 [1]. Muscle invasive bladder cancer (MIBC) is characterized by a high risk of relapse and metastasis [2]. The standard treatment consists of neoadjuvant chemotherapy followed by radical cystectomy. Nevertheless, neoadjuvant chemotherapy is a cisplatin-based schedule that is associated with significant toxicity. Some patients do not have benefit form this approach, with tumors progressing despite the administration of chemotherapy. Therefore, these patients will be receiving a toxic and unnecessary treatment, as well as delaying a potentially curative treatment, such as surgery. Unfortunately, we do not have reliable biomarkers to guide us in patient selection for these therapies. Several translational studies have aimed to identify subgroups of patients with different clinical behavior.

Choi *et al.* identified three groups of MIBC (luminal, basal and p53-like) with different response to neoadjuvant chemotherapy [3]. The Cancer Genome Atlas (TCGA) developed a molecular classification of MIBC based on RNAseq data and hierarchical cluster analysis [4]. In this work, five different groups were established: luminal-papillary (which included luminal tumors with papillary histology), luminal-infiltrated (characterized for lymphocyte infiltration), luminal, basal/squamous (also with lymphocyte infiltration) and a small neuronal group.

Seiler *et al*. associated TCGA molecular subtypes with response to neoadjuvant chemotherapy in a new cohort of patients [5]. Basal tumors appeared to benefit most from neoadjuvant chemotherapy, whereas luminal immune infiltrated tumors had a

worse prognosis. However, these findings are not compelling enough to drive clinical decisions, so further insight into the molecular biology of MIBC is needed.

Data was analyzed using three mathematical methods that have proved to be very useful in other fields. Probabilistic graphical models (PGM) can identify differences in biological process among tumors [6-9]. Mathematical classification methods, such as sparse k-means [10] and Consensus Cluster [11], previously demonstrated their utility in the establishment of tumor subtypes [6]. On the other hand, Flux Balance Analysis (FBA) is a widely used approach for modeling biochemical networks. FBA could be used to calculate tumor growth rate [12].

In this study, data from the TCGA cohort were analyzed through PGM and computational analysis to characterize MIBC at the functional level.

Material and methods

TCGA cohort: data pre-processing

TCGA RNAseq data from patients with MIBC and treated with surgery alone were used to perform this study. Patients treated with neoadjuvant therapy were initially excluded for computational analysis. For survival analysis (which relied on clinical information), subjects who had received targeted therapies or radiotherapy were excluded, as well as those with M1 disease or missing T-stage information.

Log2 of the data was calculated and genes appearing in less than 75% of the samples were discarded. Missing values were imputed using a normal distribution with Perseus software [13], as previously described [7].

Probabilistic graphical models

2,345 genes with highest variation in expression (standard deviation >2) were selected to build the PGM. The PGM method is compatible with high-throughput data with correlation as associative coefficient, as previously described [6-9]. Briefly, gene expression data was used without other *a priori* information and the analyses were done using *grapHD* package [14] and R v3.2.5 [15]. The resulting network was split into several branches and the most representative function of each branch was established by gene ontology analyses using DAVID webtool, as previously described. [16]. Functional node activities were calculated by the mean expression of the genes related to the main function assigned to each node.

Biological layer analyses

Sparse k-means [10] and Consensus Cluster Analysis (CCA) [11] were used to explore molecular groups in the TCGA MIBC data, as previously described [6]. Sparse k-means assigns a weight to each variable, based on its relevance in the sample classification. Then, a CCA using variables that were selected by the sparse k-means method was applied to define the optimum number of groups for each case. The sparse k-means-CCA workflow was performed several times to explore the presence of independent informative molecular layers. Once relevant genes for one molecular layer were identified, these genes were removed from the dataset and the sparse k-means-CCA workflow was performed again, allowing the identification of different layers of molecular information and establishing different classifications based on various molecular features. Gene ontology analyses were performed for each layer to derive functional information.

Flux Balance Analysis and flux activities

FBA was used to build a computational model that predicts tumor growth rates. COBRA Toolboox available for MATLAB [17], and the human whole metabolic reconstruction Recon2 [18] were used. The “biomass” reaction included in the Recon2 was designated as objective function. As described in previous works [8, 9], expression data was included into the model by solving GPR rules and using a modified E-flux algorithm [9, 19].

As in previous works [9], flux activities for each pathway were calculated by the sum of fluxes for all reactions involved in one pathway as defined in the Recon2. Then,

comparisons between luminal and basal groups were performed using a Mann-Whitney test.

Statistical analyses

GraphPad Prism v6 was used for statistical analyses. All p-values were two-sided and considered statistically significant under 0.05. Network analyses were done using Cytoscape software [20].

Results

Data pre-processing and patient selection

The TCGA cohort included 427 patients. Ten patients who had received neoadjuvant chemotherapy were excluded, leaving 417 subjects for subsequent analyses. Patients treated with targeted therapy or radiotherapy; M1 at diagnosis or no specified muscle-invasive type in the database were excluded from the analyses involving clinical data. Therefore, 178 patients were used for survival analyses (Sup Figure 1).

Patients and samples characteristics

Data from 178 patients included in the clinical analyses are summarized in Sup Table 1. Median of overall survival, considering a five years period of follow-up, is 1,270 days and there were 73 death events during this time.

Functional network

PGM were used to build a network, as previously described [6-8, 21] and the resulting network was functionally characterized. The network included 13 branches or functional nodes, one of them without a main relevant biological function and one with two different main biological functions: cytochrome metabolism and steroid metabolism (Figure 1).

Biological layer analyses

The sparse k-means-CCA workflow defined 16 different layers of information (Sup Table 2). The first three layers had different ontologies and were further analyzed.

Layers 4 to 13 had similar ontologies that one of the first three layers, so they were dismissed.

Layer 1: Extracellular exosome and epidermis development

Layer 1 was based on 75 genes, which were mainly related with extracellular exosome, epidermis development and sodium ion homeostasis. This layer divided patients into two different groups. Group 1.1 included 260 patients (62.35%) and was characterized by lower expression of genes included in the epidermis development and extracellular matrix nodes. Group 1.2 included 157 patients (37.64%) and showed higher expression of genes included in the epidermis development and extracellular matrix nodes (Sup Figure 2). The TCGA classification of MIBC establishes the existence of both luminal and basal groups. Interestingly, our Group 1.1 had a higher expression of KRT20, GATA3 and FOXA1 genes, all of them luminal biomarkers, whereas Group 1.2 had a higher expression of KRT5, KRT6 and KRT14 genes, all of them basal biomarkers (Figure 2 and 3). So, from now on, Group 1.1 will be called Luminal group and Group 1.2, Basal group. Luminal tumors had a trend towards better survival than basal tumors, although the difference was not statistically significant (p= 0.1210, HR= 0.6969) (Sup Figure 3).

Functional node activity for each node was calculated and compared between these two groups as previously described [6-8]. There were significant differences between luminal and basal subgroups in cytochrome metabolism, steroid metabolism, membrane, DNA binding, stem cell pathways, epidermis development, growth,

extracellular matrix, adaptive immune response, innate immune response, extracellular space, and CNS development functional node activity (Sup Figure 4).

Differences in steroid metabolism node between luminal and basal tumors led us to evaluate the expression of androgen receptor (AR) in both groups. Interestingly, Luminal tumors showed higher expression of the AR gene ($p<0.0001$, fold change = 2.669) (Figure 4).

Layer 2: Extracellular space

Layer 2 was based on 82 genes mainly related with extracellular space. Layer 2 classifieds 268 patients (64.3%) in Group 2.1 and 149 patients (35.7%) in Group 2.2. Group 2.1 was characterized by higher expression of the extracellular matrix node and lower expression of the cytochrome metabolism node. Group 2.2 had the opposite expression pattern (Sup Figure 5). Group 2.1 had better prognosis than Group 2.2(Sup Figure 6).

Layer 3: Immune

Layer 3 was based on 66 genes mainly related with inflammatory immune response. This layer divided patients into two groups. Group 3.1 included 215 patients (51.55%) and Group 3.2, 202 patients (48.44%). Group 3.1 was characterized by a high expression of innate and adaptive immune response nodes so, from now on, it will be called immune-high group. Group 3.2 was characterized by a low expression of innate and adaptive immune response nodes and will be called immune-low group (Sup Figure 7). The TCGA study used CD274 and CTLA4 to define immune infiltration in both luminal and basal tumors. In our new groups, these two immune biomarkers were

more expressed in the immune-high group (Figure 5). In addition, the immune-low group had better prognosis, although the difference was not statistically significant (Sup Figure 8). As expected, the node activities of immune nodes were higher in the immune-high group (Sup Figure 9). Both the basal and the luminal groups contained immune-high and immune-low tumors, although the basal group had less immune-high tumors (Sup Figure 10).

Layers 14 and 15

The first three layers provided distinct ontology information, but layers 4 to 13 contained redundant information. Layer 14 (translation) and layer 15 (chemical synapsis) provided no additional grouping information (Sup Figure 11).

Flux Balance Analysis growth predictions and flux activities

FBA was used to study tumor growth and compare it between the layer-defined groups. According FBA predictions, basal tumors were more proliferative than luminal tumors (Figure 6) ($p<0.0001$).

Flux activities were calculated for each metabolic pathway and compared between basal and luminal tumors to determine differential pathways as described previously [21]. Luminal tumors had a higher androgen and steroid metabolism flux activity. Differences were also detected in aminosugar metabolism, coA synthesis, galactose metabolism, glycolysis, hyaluronan acid metabolism, lysine, methionine, NAD, nucleotide savage, oxidative phosphorylation, phosphatidyl inositol, pyrimidine synthesis, R group, triacylglycerol and vitamin B6 metabolism (Sup Figure 12)

Comparison with TCGA classification

The TCGA classification mixes histological, immune and luminal-basal information, establishing three luminal groups: luminal, luminal-papillary, and luminal-infiltrated (by lymphocytes), one basal group characterized by an immune positive status, and a small group called neuronal [4]. Our classification established the immune information as an independent layer divided between luminal and basal groups, i.e., both of them had immune- high and immune-low tumors. Therefore, not all basal tumors were classified as immune positive. Additionally, the TCGA luminal-papillary group, which is defined solely by histological features, had immune-high tumors when the layer classification was applied. Percentages were similar between both classifications, although we did not identify a neuronal group (Sup Table 3).

Discussion

MIBC has a poor prognosis, with over 50% of relapses in spite of appropriate therapy [22]. Therefore, it is still necessary to characterize MIBC in order to propose new therapeutic targets. With the aim of characterize MIBC patients at functional and molecular level, PGM, layer analyses and FBA were performed in this study to provide insight into the molecular features of MIBC.

The PGM unveiled the functional structure of tumors, which has been previously described by our group [7-9]. This allows the study of gene expression data from biological and functional points of view. As an example of consistency in this regard, cytochrome P450 and UDP-glucuronosyltransferase were related to androgen receptor in the same node, and it is known that androgens modulate the expression of these enzymes [23].

Layer analyses provided different information about molecular features of the tumors, as for example, about the cellular adhesion process and about the immune status. The first layer divided MIBC tumors into Luminal and Basal. Luminal tumors presented a higher steroid metabolism node activity. Indeed, AR gene presents higher expression in Luminal tumors, suggesting the utility of AR as a possible therapeutic target. AR was previously related with bladder cancer progression [24] and *in vitro* studies showed that a siRNA against AR decreased proliferation of AR-positive bladder cancer cell lines but had no effect on AR-negative cells [25]. Therefore, patients with Luminal MIBC tumors, characterized by high expression of AR, could be candidates for therapy with AR inhibitors.

Luminal tumors had a higher flux activity of androgen and steroid metabolism pathway, which agrees with the results found in node activity. Luminal tumors also had a higher flux activity at glycolysis pathway so they may respond to drugs targeting metabolism as metformin, which has been shown to reduce growth in bladder cancer cells [26].

On the other hand, FBA predicted that, as it has been seen in breast cancer [8], basal tumors are more proliferative than luminal tumors. It is established that basal breast cancer tumors have a good response to chemotherapy because they are more proliferative [27, 28]. Based on the FBA results, the previous knowledge in basal breast tumors, and taking into account that this cohort is chemotherapy-naïve, basal patients may be good responders to chemotherapy as it was previously suggested by Seiler *et al.* [5]. Proliferation has been determined in other tumor types through gene expression, but data in bladder carcinoma are scarce. FBA allows not only to study proliferation but also to compare metabolic pathways.

The third layer identified an immune high expression group with high expression of CTLA4 and CD274, which may be a group of patients candidates of receiving immunotherapy, given that CD274, also known as PD-L1, and CTLA4 are the basis of current immunotherapy [29]. Interestingly, immune high tumors had a worse prognosis than immune low tumors, according with Seiler *et al.* results, which suggested that luminal immune infiltrated tumors had a worse prognosis [5].

Percentage distribution between luminal and basal tumors was comparable in the TCGA classification and the layer classification. However, the TCGA classification mixes

immune, histological and molecular information. Our approach, on the contrary, establishes two independent informative layers: molecular and immune; and it rendered some interesting findings that complement the TCGA classification: for instance, 10% of basal tumors had an immune-low status, whereas in the TCGA classification all basal tumors had an immune-positive status. With the arrival of immunotherapies to the clinic, it could be useful to characterize the immune status of tumors and establish groups with differential immune features to drive treatment decisions.

The study has some limitations. Our findings should be validated in an independent cohort. Publications including expression and clinical data are scarce, so validation would rather be performed in a prospective study. On the other hand, the existence of a neuronal group could not be confirmed. As neuronal group accounted for a minority of cases in the TCGA study, maybe we should have analyzed a larger population. Finally, although our results suggest that some drugs may work better in specific groups, this should be prospectively validated. Response to chemotherapy, for instance, has been related to multiple factors and the proliferation profile may not be enough to identify responders.

Conclusions

Computational analyses found different levels of information in gene expression data from MIBC: one of these levels refers to immune features, whereas the other corresponds to the previous classification into luminal/basal subgroups. Our classification may have therapeutic implications for the treatment of MIBC.

Take Home Message: We used computational analyses in a muscle-invasive bladder cancer cohort and we defined independent molecular and immune features in these tumors that allow us to suggest therapeutic targets.

Acknowledgments

This study was supported by Instituto de Salud Carlos III, Spanish Economy and Competitiveness Ministry, Spain and co-funded by the FEDER program, “Una forma de hacer Europa” (PI15/01310). LT-F is supported by Spanish Economy and Competitiveness Ministry (DI-15-07614). GP-V is supported by Conserjería de Educación, Juventud y Deporte of Comunidad de Madrid (IND2017/BMD7783). The funders had no role in the study design, data collection and analysis, decision to publish or preparation of the manuscript.

Disclosures

JAFV, EE and AG-P are shareholders in Biomedica Molecular Medicine SL. LT-F and GP-V are employees of Biomedica Molecular Medicine SL. The other authors declare that they have no competing interests.

Bibliography

[1] Siegel RL, Miller KD, Jemal A. Cancer statistics, 2018. CA Cancer J Clin. 2018;68:7-30.
[2] Shah JB, McConkey DJ, Dinney CP. New strategies in muscle-invasive bladder cancer: on the road to personalized medicine. Clin Cancer Res. 2011;17:2608-12.
[3] Choi W, Porten S, Kim S, Willis D, Plimack ER, Hoffman-Censits J, et al. Identification of distinct basal and luminal subtypes of muscle-invasive bladder cancer with different sensitivities to frontline chemotherapy. Cancer Cell. 2014;25:152-65.
[4] Robertson AG, Kim J, Al-Ahmadie H, Bellmunt J, Guo G, Cherniack AD, et al. Comprehensive Molecular Characterization of Muscle-Invasive Bladder Cancer. Cell. 2017;171:540-56.e25.
[5] Seiler R, Ashab HAD, Erho N, van Rhijn BWG, Winters B, Douglas J, et al. Impact of Molecular Subtypes in Muscle-invasive Bladder Cancer on Predicting Response and Survival after Neoadjuvant Chemotherapy. Eur Urol. 2017;72:544-54.
[6] de Velasco G, Trilla-Fuertes L, Gamez-Pozo A, Urbanowicz M, Ruiz-Ares G, Sepúlveda JM, et al. Urothelial cancer proteomics provides both prognostic and functional information. Sci Rep. 2017;7:15819.
[7] Gámez-Pozo A, Berges-Soria J, Arevalillo JM, Nanni P, López-Vacas R, Navarro H, et al. Combined label-free quantitative proteomics and microRNA expression analysis of breast cancer unravel molecular differences with clinical implications. Cancer Res; 2015. p. 2243-53.

[8] Gámez-Pozo A, Trilla-Fuertes L, Berges-Soria J, Selevsek N, López-Vacas R, Díaz-Almirón M, et al. Functional proteomics outlines the complexity of breast cancer molecular subtypes. Scientific Reports. 2017;7:10100.
[9] Trilla-Fuertes L, Gamez-Pozo A, M Arevalillo J, Diaz-Almiron M, Prado-Vazquez G, Zapater-Moros A, et al. Molecular characterization of breast cancer cell response to metabolic drugs. Oncotarget2018.
[10] Witten DM, Tibshirani R. A framework for feature selection in clustering. J Am Stat Assoc. 2010;105:713-26.
[11] Monti S, Tamayo P, Mesirov J, Golub T. Consensus Clustering: A Resampling-Based Method for Class Discovery and Visualization of Gene Expression Microarray Data. Machine learning. 2003;52:91-118.
[12] Orth J, Thiele I, Palsson B. What is flux balance analysis? : Nat Biotechnol; 2010. p. 245-8.
[13] Tyanova S, Temu T, Sinitcyn P, Carlson A, Hein MY, Geiger T, et al. The Perseus computational platform for comprehensive analysis of (prote)omics data. Nat Methods. 2016;13:731-40.
[14] Abreu G, Edwards D, Labouriau R. High-Dimensional Graphical Model Search with the gRapHD R Package Journal of Statistical Software2010. p. 1-18.
[15] R Core Team. R: A language and environment for statistical computing. Vienna,Austria. R Foundation for Stattistical Computing, 2013.
[16] Huang dW, Sherman BT, Lempicki RA. Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. Nat Protoc. 2009;4:44-57.
[17] Schellenberger J, Que R, Fleming R, Thiele I, Orth J, Feist A, et al. Quantitative prediction of cellular metabolism with constraint-based models: the COBRA Toolbox v2.0. Nature Protocols; 2011. p. 1290-307.
[18] Thiele I, Swainston N, Fleming RM, Hoppe A, Sahoo S, Aurich MK, et al. A community-driven global reconstruction of human metabolism. Nat Biotechnol. 2013;31:419-25.
[19] Colijn C, Brandes A, Zucker J, Lun D, Weiner B, Farhat M, et al. Interpreting expression data with metabolic flux models: Predicting Mycobacterium tuberculosis mycolic acid production. PLOS Comput Bio; 2009.
[20] Shannon P, Markiel A, Ozier O, Baliga NS, Wang JT, Ramage D, et al. Cytoscape: a software environment for integrated models of biomolecular interaction networks. Genome Res. 2003;13:2498-504.
[21] Trilla-Fuertes L, Gámez-Pozo A, Arevalillo JM, Díaz-Almirón M, Prado-Vázquez G, Zapater-Moros A, et al. Molecular characterization of breast cancer cell response to metabolic drugs. Oncotarget. 2018;9:9645-60.
[22] Bellmunt J, Orsola A, Leow JJ, Wiegel T, De Santis M, Horwich A, et al. Bladder cancer: ESMO Practice Guidelines for diagnosis, treatment and follow-up. Ann Oncol. 2014;25 Suppl 3:iii40-8.
[23] Imaoka S, Yoneda Y, Sugimoto T, Ikemoto S, Hiroi T, Yamamoto K, et al. Androgen regulation of CYP4B1 responsible for mutagenic activation of bladder carcinogens in the rat bladder: detection of CYP4B1 mRNA by competitive reverse transcription-polymerase chain reaction. Cancer Lett. 2001;166:119-23.
[24] Jitao W, Jinchen H, Qingzuo L, Li C, Lei S, Jianming W, et al. Androgen receptor inducing bladder cancer progression by promoting an epithelial-mesenchymal transition. Andrologia. 2014;46:1128-33.
[25] Miyamoto H, Yang Z, Chen YT, Ishiguro H, Uemura H, Kubota Y, et al. Promotion of bladder cancer development and progression by androgen receptor signals. J Natl Cancer Inst. 2007;99:558-68.
[26] Zhang T, Guo P, Zhang Y, Xiong H, Yu X, Xu S, et al. The antidiabetic drug metformin inhibits the proliferation of bladder cancer cells in vitro and in vivo. Int J Mol Sci. 2013;14:24603-18.

[27] De Giorgi U, Rosti G, Frassineti L, Kopf B, Giovannini N, Zumaglini F, et al. High-dose chemotherapy for triple negative breast cancer. Ann Oncol. England2007. p. 202-3.
[28] Liedtke C, Mazouni C, Hess KR, Andre F, Tordai A, Mejia JA, et al. Response to neoadjuvant therapy and long-term survival in patients with triple-negative breast cancer. J Clin Oncol. 2008;26:1275-81.
[29] Pico de Coaña Y, Choudhury A, Kiessling R. Checkpoint blockade for cancer therapy: revitalizing a suppressed immune system. Trends Mol Med. 2015;21:482-91.

Figure legends:

Figure 1: Probabilistic graphical model graph showing the network functional structure.

Each node is named as its gene ontology main function identified.

Figure 2: Expression of Basal biomarkers in Group 1.1 and Group 1.2.

Figure 3: Expression of Luminal biomarkers in Group 1.1 and Group 1.2.

Figure 4: Androgen receptor expression in Luminal and Basal group.

Figure 5: Expression of immune biomarkers in our immune groups.

Figure 6: Tumor growth rate predicted by FBA in luminal and basal tumors.

Supplementary table legends:

Sup Table 1: Clinical patients' characteristics.

Sup Table 2: Main gene ontology term defined for each sixteen layers obtained by the sparse k-means-CCA workflow.

Sup Table 3: Percentages of patients assigned to each group in TCGA and layer classification.

Supplementary figure legends:

Sup Figure 1: Flowchart for patient selection.

Sup Figure 2: PGM's graph heatmap showing differences between Group 1.1 (Luminal) and Group 1.2 (Basal). Green= underexpressed, Red= overexpressed.

Sup Figure 3: Kaplan Meier analysis comparing Luminal and Basal MIBC tumors clinical evolution.

Sup Figure 4: Functional node activities comparison between Luminal and Basal group.

Sup Figure 5: PGM's graph heatmap showing differences between Group 2.1 and Group 2.2. Green= underexpressed, Red= overexpressed.

Sup Figure 6: Kaplan Meier analysis comparing Group 2.1 and Group 2.2 clinical evolution.

Sup Figure 7: Heatmap showing differences between Group 3.1 (immune-high) and Group 3.2 (immune-low). Green= underexpressed, Red= overexpressed.

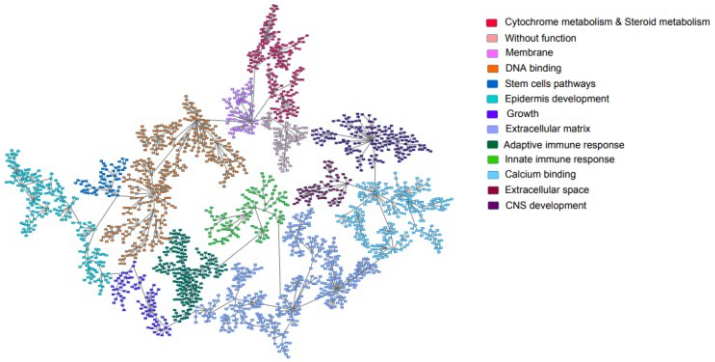
Sup Figure 8: Survival curves obtained for immune groups.

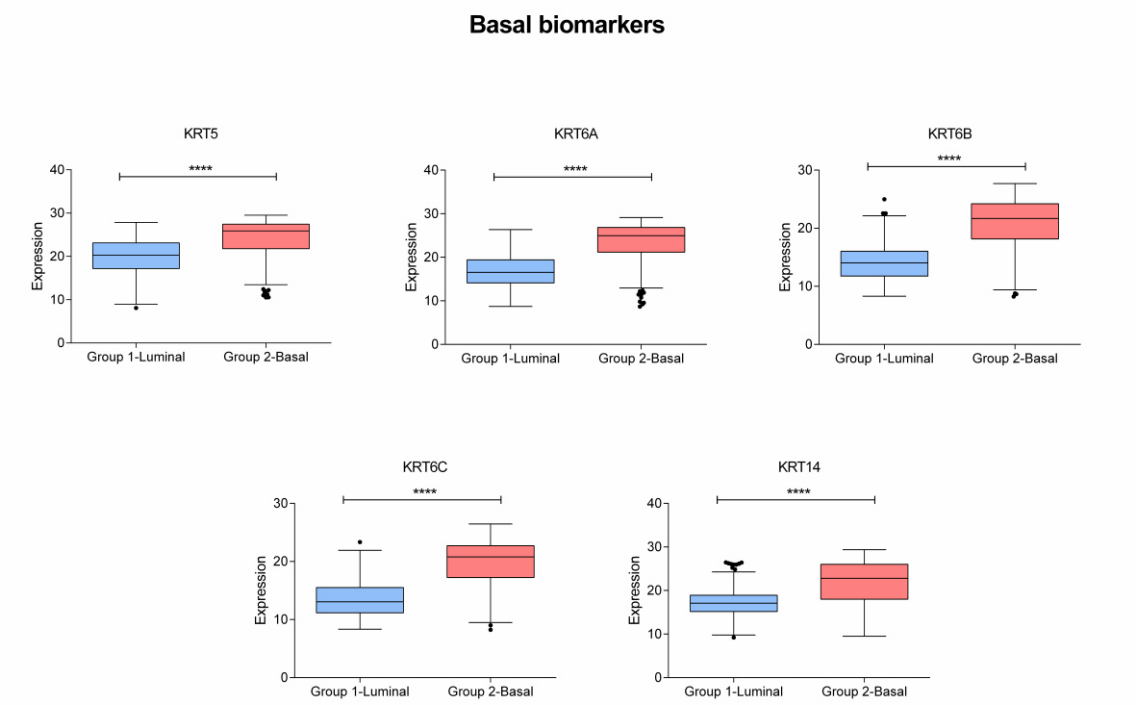
Sup Figure 9: Immune node activities in immune groups.

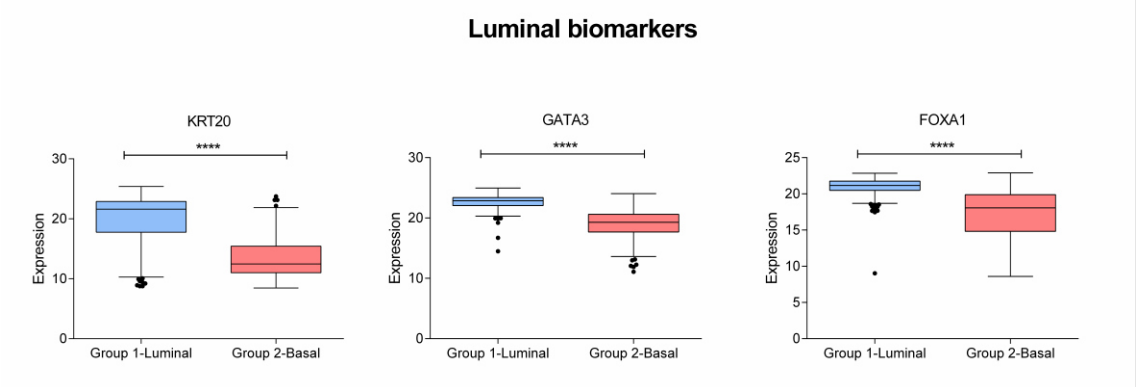
Sup Figure 10: Concordance between classification of layer 1, which divided tumors into Luminal and Basal, and layer 3, which divided tumors into immune-high and immune-low group.

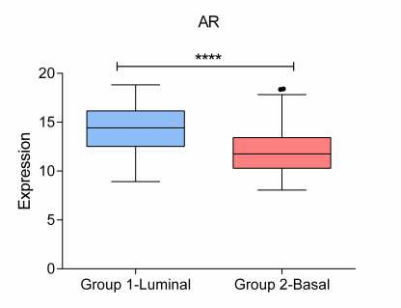
Sup Figure 11: Heatmap showing differences between groups defined in layers 14 and 15. Green= underexpressed, Red= overexpressed.

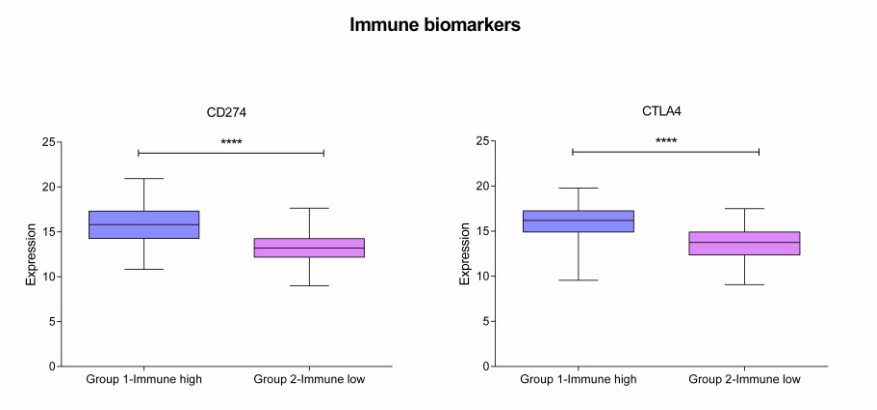
Sup Figure 12: Flux activities of luminal and basal groups.

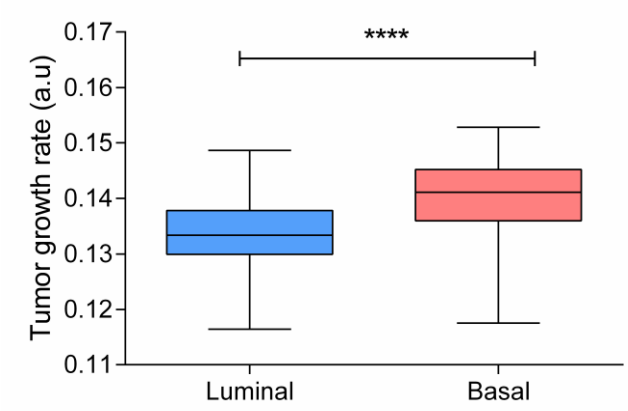












Bayesian Networks established functional differences between breast cancer subtypes

Lucía Trilla-Fuertes[1¶], Andrea Zapater-Moros[2¶], Angelo Gámez-Pozo[1,2], Jorge M Arevalillo[3], Guillermo Prado-Vázquez[1], Mariana Díaz-Almirón[4], María Ferrer-Gómez[2], Rocío López-Vacas[2], Hilario Navarro[3], Enrique Espinosa[5,6], Paloma Maín[7], and Juan Ángel Fresno Vara[1,2,7*]

[1]Biomedica Molecular Medicine SL, Madrid, Spain

[2]Molecular Oncology & Pathology Lab, Institute of Medical and Molecular Genetics-INGEMM, La Paz University Hospital-IdiPAZ, Madrid, Spain

[3]Operational Research and Numerical Analysis, National Distance Education University (UNED), Spain

[4]Biostatistics Unit, La Paz University Hospital-IdiPAZ, Madrid, Spain

[5]Medical Oncology Service, La Paz University Hospital-IdiPAZ, Madrid, Spain

[6] Biomedical Research Networking Center on Oncology-CIBERONC, ISCIII, Madrid, Spain

[7]Department of Statistics and Operations Research, Faculty of Mathematics, Complutense University of Madrid, Madrid, Spain

[¶] These authors contributed equally to this work.

*Corresponding Author:

E-mail: juanangel.fresno@salud.madrid.org (JAFV)

Abstract

Breast cancer is a heterogeneous disease. In clinical practice, tumors are classified as hormonal receptor positive, Her2 positive and triple negative. In previous studies, our group defined a new hormonal receptor-positive subgroup, the TN-like subtype, which has a prognosis and a molecular profile more similar to triple negative tumors. In this study, proteomics and Bayesian networks were used to characterize protein relationships in 106 breast tumor samples. Components obtained by these methods had a clear functional structure. The analysis of these components suggested differences in processes such as metastasis or proliferation between breast cancer subtypes, including our new TN-like subtype. In addition, one of the components, mainly related to metastasis, had prognostic value in this cohort. Functional approaches enable to build hypotheses on regulatory mechanisms and establish new relationships among proteins in the breast cancer context.

Author Summary:

Breast cancer classification in clinical practice, as defined by three biomarkers (estrogen receptor, progesterone receptor and HER2), is categorized into hormone receptor positive, HER2+ or triple negative breast cancer (TNBC). Our group recently reported a new ER+ subtype with molecular characteristics and a prognosis similar to TNBC. In this study, we propose a mathematical method, the Bayesian networks, as a useful tool to study protein interactions and differential biological processes in breast cancer subtypes, characterizing differences in relevant processes such as proliferation or metastasis, and associating them with patient prognosis.

Introduction

Breast cancer is one of the most prevalent cancers in the world [1]. In clinical practice, breast cancer is classified, according to the expression of hormonal receptors (estrogen or progesterone) and Her2, into hormonal receptor positive (ER+), human epidermal growth factor receptor 2 positive (HER2+) and triple negative breast cancer (TNBC). In previous studies, our group defined a new ER+ molecular subgroup, named TN-like, with a molecular profile and a prognosis more similar to TNBC tumors [2]. The remaining ER+ tumors were considered as ER-true. We also found significant molecular differences among breast cancer subtypes. For instance, differences related to glucose metabolism were described between ER-true, TN-like and TNBC tumors [2, 3].

Proteomics provides useful information about biological process effectors and can quantify thousands of proteins. Undirected probabilistic graphical models (PGMs) based on a Bayesian approach allow us to characterize differences between tumor samples at a functional level [25]. In this study, we explored the utility of Bayesian networks in the molecular characterization of breast cancer. The main feature of targeted Bayesian networks is that they provide a hierarchical structure and targeted relationships between proteins.

In this study, we used proteomics and Bayesian networks to characterize protein relationships in a cohort of breast cancer tumor samples. These networks maintain a functional structure, and it is possible to use this information to build prognostic signatures. This approach also reflects previously described interactions, and it could be used to propose new hypotheses and mechanisms of regulation of these proteins.

Materials and methods

Ethics statement

Informed consent was obtained from the study participants. Approval for the study was obtained from the ethics committees of Hospital Doce de Octubre and Hospital Universitario La Paz.

Samples

A total 106 FFPE samples from patients with breast cancer were retrieved from the I+12 Biobank and from the IdiPAZ Biobank, both integrated within the Spanish Hospital Biobank Network. The histopathological characteristics were reviewed by a pathologist to confirm tumor content. Samples had to comprise no less than half of tumor cells. These samples had been used in previous studies [2, 3, 6].

Protein preparation

Proteins were extracted from the FFPE samples as previously described [7]. Briefly, FFPE sections were deparaffinized in xylene and washed twice with absolute ethanol. Protein extracts from the FFPE samples were set up in 2% sodium dodecyl sulfate (SDS) buffer using a protocol based on heat-induced antigen retrieval. Protein concentration was quantified using the MicroBCA Protein Assay Kit (Pierce-Thermo Scientific). Protein extracts (10 µg) were processed with trypsin (1:50) and SDS was removed from digested lysates using Detergent Removal Spin Columns (Pierce). Peptide samples were additionally desalted using ZipTips (Millipore), dried, and resolubilized in 15 µL of a 0.1% formic acid and 3% acetonitrile solution before the MS experiments.

Label-free proteomics

Samples were analyzed on a LTQ Orbitrap Velos hybrid mass spectrometer (Thermo Fisher Scientific, Bremen, Germany) coupled to a NanoLC Ultra system (Eksigent Technologies, Dublin, CA, USA) as described previously [2, 3]. Briefly, after separation, peptides were eluted with a gradient of 5% to 30% acetonitrile in 95 minutes. The mass spectrometer was operated in data-dependent mode (DDA), followed by collision-induced dissociation fragmentation on the 20 most intense signals per cycle. The acquired raw MS data were processed by MaxQuant (version 1.2.7.4) [8], followed by protein identification using the integrated Andromeda search engine [9]. Briefly, spectra were searched against a forward UniProtKB/Swiss-Prot human database, concatenated to a reversed decoyed FASTA database (NCBI taxonomy ID 9606, release date 2011-12-13). The maximum false discovery rate (FDR) was set to 1% for peptides and 5% for proteins. Label-free quantification was calculated on the basis of the normalized intensities. Quantifiable proteins were defined as those detected in at least 75% of samples in at least one type of sample (either ER+ or TNBC samples) showing two or more unique peptides. Only quantifiable proteins were considered for subsequent analyses. Protein expression data were log2 transformed and missing values were replaced using data imputation for label-free data, as explained in Deeb et al. [10], using default values. Finally, protein expression values were z-score transformed. Batch effects were estimated and corrected using ComBat [11]. All the mass spectrometry raw data files acquired in this study can be downloaded from Chorus (http://chorusproject.org) under the project name Breast Cancer Proteomics.

Network construction

PGMs are graph-based representations of joint probability distributions in which nodes represent random variables, and edges (directed or undirected) represent stochastic

dependencies among the variables. In particular, we have used a type of PGM called Bayesian networks (BNs) [12]. With these models, the dependences between the variables in our data are specified by a directed acyclic graph (DAG).

First, we found the BN that best explained our data [13]. Although there are various algorithms to create a DAG from data, we selected the well-known PC algorithm, a constraint-based structure learning algorithm [14] based on conditional independence tests. The PC algorithm was shown to be consistent in some high-dimensional settings [15] and it has some other statistical properties that make it very useful, such as high computational efficiency and consistency for very high-dimensional, sparse DAGs [16]; robustness [17]; and high-dimensional consistency that carries over to a broader class of Gaussian copula or nonparanormal models when using rank-based measures of correlation [18]. Thus, our data are represented by a large graph that can be partitioned into several connected components. Then, we focused on finding suitable subgraphs to give us a much clearer understanding of the interrelations therein. All these procedures are implemented in R [19], within packages *pcalg* [15] and *graph* [20]. We used protein expression data without other *a priori* information.

With the aim of comparing and completing the information provided by the BN, the Genes2FANS (G2F) tool, developed by the Ma'ayan group was used [21]. This software provides protein-protein interactions (PPI) information based on experimental evidence. A PPI network using G2F was built for each component, and finally, both the DAG and PPI networks were merged.

Gene ontology analyses

Protein to Gene Symbol conversion was performed using Uniprot (www.uniprot.org) and DAVID (www.david.ncifcrf.gov) [22]. The gene ontology analysis was also performed in DAVID, selecting only the "*Homo sapiens*" background and the GOTERM-FAT, Biocarta and KEGG

133 databases. In terms of small connected components, a literature search was performed so as
134 to establish main component functions.

135 Component activity measurements

136 Component activities were calculated as previously described [3, 4]. Briefly, activity
137 measurement was calculated by the mean expression of all the proteins of each component
138 related to the established major component function.

139 Prognostic model development

140 Prognostic signatures were developed using R v3.2.4 and BRB Array Tools, developed by
141 Richard Simon and the BRB Array Tools Development Team [23]. As described in previous
142 studies [6], we identified the component activity measurements related to the distant
143 metastasis-free survival (DMFS) based on their p-values. We then used the component activity
144 measurements with a p-value less than 0.05 to build a prognostic signature. The cut-off was
145 established *a priori* in order to avoid overfitting the predictive signature to our population;
146 thus, the cut-off was based on the relapse proportions in the cohort.

147 Statistical analyses

148 The statistical analyses were performed using GraphPad Prism v6. The network analyses were
149 performed using Cytoscape software [24]. P-values less than 0.05 were considered statistically
150 significant.

Results

Patient characteristics

The clinical characteristics of this patient cohort have been previously described [2, 3, 6]. Briefly, 106 patients were enrolled in the study; all had node positive disease, were Her2 negative and had received adjuvant chemotherapy and hormonal therapy in the case of ER+ tumors. Among the ER+ tumors, 50 patients were characterized as ER-true and 21 were defined as TN-like (S1 Table) [2].

Mass-spectrometry analysis

Proteomics analyses from these samples have been previously described [3]. In summary, 102 formalin-fixed paraffin-embedded (FFPE) samples had sufficient protein to perform the mass-spectrometry (MS) analyses. After MS workflow, 96 samples provided useful protein expression data. After quality criteria were applied, 1095 proteins were found to present at least two unique peptides and detectable expression in at least 75% of the samples in at least one type of sample (either ER+ or TNBC).

Directed networks

Using proteomics data, DAGs were created. Altogether, it was possible to establish 536 edges, of which 414 were directed and 122 were undirected. These edges formed 377 components formed by different numbers of nodes or proteins. An overview of the number of nodes (proteins) included in each component is provided in Table 1.

Number of nodes	1	2	3	4	5	6	7	8	9	10	11	13	14	19	21	26	27	34	36
Number of components	229	70	24	12	15	2	6	3	4	2	1	2	1	1	1	1	1	1	1

171 **Table 1: Characteristics of the components obtained from DAGs.** Number of nodes = number
172 of proteins contained in each component; number of components = directed components
173 obtained.

174 We characterized components from the DAG analysis. Components including fewer than 9
175 nodes were dismissed because they conveyed little information. All the components were
176 named with the number of nodes included in the DAG analysis, and the information was
177 completed with PPIs based on experimental evidence obtained from the G2F database tool
178 [21]. For example, component 14 includes 19 nodes, 14 nodes defined from our Bayesian
179 analysis and 5 extra nodes added by G2F.

180 Afterward, components were investigated for biological function. For components with fewer
181 than 27 nodes provided by the DAG analysis, the functional analysis was performed by
182 bibliographical review. Thus, the main biological function of component 14 is metabolism.
183 Regarding components that had more than 27 nodes provided by the DAG analysis, gene
184 ontology analyses were performed once G2F information was complete. Characteristics of all
185 components are supplied in Table 2 and S1 File.

Component	Main function	Total number of nodes
Component 36	Extracellular exosome	65 nodes
Component 34	RNA processing and mTOR pathway	61 nodes
Component 19	Proliferation	26 nodes
Component 21	Proliferation & metastasis	34 nodes
Component 27	Metastasis	34 nodes
Component 14	Metabolism	19 nodes
Component 13	Proteasome	19 nodes
Component 11	No main function assigned	14 nodes
Component 10b	Proliferation	19 nodes
Component 9a	Growth	15 nodes
Component 9b	Proliferation & metastasis	22 nodes
Component 9c	Glycolysis	12 nodes
Component 9d	Transcription & ribosomes	20 nodes

186 **Table 2: Features of components obtained by DAG and G2F analysis.**

Component activity measurements

Component activities were calculated for each node. There were significant differences between ER-true, TN-like and TNBC tumors in the component activity for component 34: mRNA processing and mTOR; component 36: extracellular exosome; component 21: proliferation and metastasis; component 27: metastasis; component 9a: growth; component 9b: proliferation and metastasis; component 9c: glycolysis; component 10b: metastasis; component 13: proteasome; and component 9d: transcription (Fig. 1).

Fig 1: Component activity measurements for ER-true, TN-like and TNBC.

Component 10b: metastasis

Component 10b activity showed prognostic value in our series, splitting our population into a high- and a low-risk group, and could be used to make a DMFS predictor (p=0.007; HR= 0.29; cut-off 40% low risk :60% high risk) (Fig. 2). Dividing by molecular subtype, this prognostic signature also split the population into low- and high-risk groups, although it was not statistically significant (Fig. 3).

Fig 2: Component 10b activity prognostic value in the entire cohort.

Fig 3: Component 10b activity prognostic value by subtype.

Component 10b contains 10 proteins. Four out of nine edges (DYNLRB1-HSPH1; HSPH1-CFL1; HBB-UCHL5; and UCHL5-CFL1) have been previously described in the G2F database [21]. G2F then added two more proteins (HSPH1 and UCHL5) to this component. Most proteins included in this component were related to metastatic processes [25-29], whereas PBDC1 had no associated function. On the other hand, HIST2H2BF is a histone, and HIST2H3PS2 is a histone pseudogene, showing a directed relationship (Fig. 4).

209 **Fig 4: Component 10b merged with PPI information provided by Genes2FANS.** Red arrows
210 indicate relationships from the DAGs; green lines indicate relationships from the Genes2FANS
211 database. Orange nodes are common nodes between these two approaches; purple nodes are
212 only from the DAGs and blue nodes are only from the Genes2FANS.

213

Discussion

In this study, we used proteomics and DAGs to characterize relationships between proteins in breast cancer tumor samples. Unlike other approaches, such as G2F [21], our DAG method supplies directed relationships between proteins and a hierarchical structure. Traditionally, PPI networks are based on relationships described in the literature. However, we built a directed network, i.e., a graph formed by edges with a direction, using protein expression data without other *a priori* information; thus, it was possible to propose new hypotheses about protein interactions. We used PGMs because they offer a way to relate many random variables to a complex dependency structure.

Arrows in directed networks indicate causality between two proteins; i.e., if proteins A and B are connected and protein A changes its expression value, protein B changes its expression value as well. This approach enables to make hypotheses about causal relationships between proteins and proposes a hierarchical structure. The G2F relationships supplied additional information to our directed networks; thus, they served as a validation of the network coherence. In some cases, an experimental relationship between two proteins connected in the directed network had been described. In component 10b, for instance, a common HSPH1 nexus for CFL1 and DYNRBL1 had been found [21]. In addition, the DAG analysis established a relationship between PTRF and PRDXCDBP in component 27, which had been widely reported [30].

We demonstrated in previous studies that non-directed graphs provided functional information [2-5]. Interestingly, a functional structure also appeared in the type of network used in this study. Component activities suggested differences in functions such as metastasis, proliferation, proteasome or glycolysis. We have previously described differences in

proteasome and glycolysis between ER-true and TN-like subtypes using non-directed networks [2].

On the other hand, the role that the actin cytoskeleton plays and its regulation in directional migration and metastasis it is widely known; thus, it is not surprising that the proteins that comprise component 10b had some relationship to this biological function and had prognostic value. Using, for example, component 10b activity, based on the proteins related to metastasis, it is possible to split our population into groups at low and a high risk of relapse; interestingly, this prognostic value was also shown in the analysis by subtype. In previous studies, we have used functional node activities from non-directed networks to develop prognostic predictors [4, 5]. This approach has also been validated in directed networks.

Component 10b presents three proteins showing influence over others: HIST2H2BF, DYNLRB1 and RHOG. Dynein light chain roadblock-type 1 (DYNRBL1), also known as km23-1, is a component of the cytoplasmic dynein 1 complex. In colon cells, its depletion can block events known to be involved in cell invasion and tumor metastasis, and it is also an actin cytoskeletal linker critical for the dynamic regulation of cell motility and invasion because it induces a highly organized actin stress fiber network [31]. DYNRBL1 regulates Ras-homolog family member A (RhoA) and motility-associated actin modulating proteins, such as cofilin1 (CFL1) and coronin [32], suggesting that DYNRBL1 could represent a novel target for anti-metastatic therapy [26].

On the other hand, Ras-homolog family member G (RhoG) encodes a member of the Rho family of small GTPases. Constitutively active RhoG induces morphological and cytoskeletal changes similar to those induced by the combination of active Rac and Cdc42 working upstream from them. Rac is activated at the leading edge of motile cells and induces the formation of actin-rich lamellipodia protrusions, which serves as a major driving force for cancer cells. The major downstream proteins for Rac are the WAVE family proteins, the

activators of the Arp2/3 complex. In addition, RhoG induces translocation of the Dock4-ELMO complex to the plasma membrane and enhances Rac1 activation, which promotes migration [33]. Moreover, RNA interference-mediated knockdown of RhoG in HeLa cells reduced cell migration in Transwell and scratch-wound migration assays [28].

HIST2H2BF has been proposed as a pancreatic ductal adenocarcinoma biomarker [34], however, there is no available information about its role in metastasis or breast cancer.

In component 10b, both DYNLRB1 and RHoG expression modulates CFL1, which is an actin-modulating protein related to filamentous F-actin and G-actin polymerization. Increased phosphorylation of this protein by LIM kinase aids in Rho-induced reorganization of the actin cytoskeleton [35]. CFL1 is an actin-severing protein that creates free barbed ends in the actin-severing process. Arp2/3 binding to these barbed ends allows the elongation of new actin filaments. The synergy between CFL1 severing activity and Arp2/3-generated dendritic nucleation results in the formation of stable invadopods and directional migration, linking CFL1 and ARP2/3, through RhoG, to tumor cell invasion [32, 36]. The CFL1 pathway was previously related to metastatic processes in breast cancer [35].

Furthermore, DYNRBL1 is required for TGFb1 secretion. The TGFb pathway apparently plays an important role in cell migration in this sense, because TGFbRII signaling regulates Rho GTPase degradation and actin dynamics. Moreover, the TGFb pathway induces both NET1 expression, which promotes actin polymerization, and tropomyosin expression related to cell motility, during the epithelial–mesenchymal transition process [32]. Additionally, TGFb production by macrophages or dendritic cells that have engulfed apoptotic cells can promote the generation of inducible regulatory T cells that play a known protumoral role [33].

Therefore, the relationships between CFL1, RHOG and DYNLRB1 are well-established. Interestingly, the DAG adds the decorin (DCN) to these edges. DCN is a small leucine-rich proteoglycan that promotes matrix organization by decreasing collagen uptake/degradation,

which constitutes a physical barrier against migration/motility of cancerous cells. The alteration in the matrix stiffness leads to differential integrin activation and changes in cytoskeletal organization by Rac, affecting cell motility and invasiveness [37]. Moreover, DCN acts as a matrikine, whose interaction with CXC chemokine receptor 4 (CXCR4) impairs its binding with stromal cell-derived factor-1a (SDF-1a), preventing directional migration [37]. In breast cancer tumors, high DCN expression in stroma correlated with lower tumor grade, low Ki67 levels and ER positivity. On the other hand, high expression of DCN in the malignant epithelium correlated with lymph node positivity, a higher number of positive lymph nodes and HER2-positive status compared with patients with low DCN expression [38].

Finally, lysyl-tRNA synthetase (KARS), also known as KRS, was recently shown to induce cancer cell migration through its interaction with the 67-kDa laminin receptor (67LR), which binds laminin on extracellular matrix. The interaction of KRS with 67LR enhances the membrane stability of 67LR, which in turn results in an increase in laminin-dependent cell migration in metastasis [35]. KARS causes incomplete epithelial-mesenchymal transition and ineffective cell‑extracellular matrix adhesion for migration [39].

We used mathematical DAG analyses and applied them to proteomics data of breast cancer tumors in order to infer causal relationships between proteins. However it is possible that DAG analyses suffer small variations by the introduction of a new variable, this technique seems useful, on the one hand, to build a descriptive model of a dataset and, on the other hand, to associate proteins with the same biological function. In addition, this method supplied some known relationships but also proposed new ones, and it associated proteins having a similar function. Therefore, this method appears to be a good approach for proposing new hypotheses about mechanisms of action. Moreover, it was possible to associate the results obtained by DAG analysis with prognoses and to build a prognostic signature. As far as we

311 know, this is the first time that this type of analysis has been applied to clinical data and is
312 associated with clinical outcome.
313 Our study has some limitations. Proteomics provides complementary information to other
314 techniques such as genomics. However, an improvement in the number of detected proteins is
315 still necessary. Validation at the cellular level is also needed.

316 In summary, in this study we used proteomics and Bayesian networks to characterize
317 relationships between proteins in breast cancer tumors. This approach reflected some
318 previously described interactions, and it could be used to propose new hypotheses and
319 mechanisms of action.

Funding statement

This study was supported by the Instituto de Salud Carlos III, Spanish Economy and Competitiveness Ministry, Spain, and was co-funded by the FEDER program, “Una forma de hacer Europa” (PI15/01310). LT-F is supported by the Spanish Economy and Competitiveness Ministry (DI-15-07614). GP-V is supported by Conserjería de Educación, Juventud y Deporte of Comunidad de Madrid (IND2017/BMD7783). Biomedica Molecular medicine SL provided support in the form of salaries for the authors LT-F, GP-V and AG-P, but did not have any additional role in the study design, data collection and analysis, decision to publish or preparation of the manuscript. The specific roles of these authors are articulated in the author contributions section. Other funders had no role in the study design, data collection and analysis, decision to publish or preparation of the manuscript.

Author contributions

All the authors directly participated in the preparation of this manuscript and have approved the final version submitted. RL-V prepared the proteomics samples. JMA, MD-A, HN and PM contributed the directed graphical models. LT-F, AZ-M, AG-P, G-PV and MF-G performed the statistical analyses, the directed graphical model interpretation and the literature review. LT-F, AG-P, JAFV and EE conceived the study and participated in its design and interpretation. LT-F and AZ-M drafted the manuscript. AG-P, JAFV and EE supported the manuscript drafting. AG-P and JAFV coordinated the study.

Competing interests

JAFV, EE and AG-P are shareholders in Biomedica Molecular Medicine SL. AG-P, LT-F and GP-V are employees of Biomedica Molecular Medicine SL. This does not alter our adherence to PLOS ONE policies on sharing data and materials. The other authors declare no competing interests.

Data Availability Statement

All the mass spectrometry raw data files acquired in this study can be downloaded from Chorus (http://chorusproject.org) under the project name Breast Cancer Proteomics.

References

1. Ferlay J, Soerjomataram I, Dikshit R, Eser S, Mathers C, Rebelo M, et al. Cancer incidence and mortality worldwide: sources, methods and major patterns in GLOBOCAN 2012. Int J Cancer. 2015;136(5):E359-86. Epub 2014/10/09. doi: 10.1002/ijc.29210. PubMed PMID: 25220842.
2. Gámez-Pozo A, Trilla-Fuertes L, Berges-Soria J, Selevsek N, López-Vacas R, Díaz-Almirón M, et al. Functional proteomics outlines the complexity of breast cancer molecular subtypes. Scientific Reports. 2017;7(1):10100. doi: 10.1038/s41598-017-10493-w.
3. Gámez-Pozo A, Berges-Soria J, Arevalillo JM, Nanni P, López-Vacas R, Navarro H, et al. Combined label-free quantitative proteomics and microRNA expression analysis of breast cancer unravel molecular differences with clinical implications. Cancer Res; 2015. p. 2243-53.
4. de Velasco G, Trilla-Fuertes L, Gamez-Pozo A, Urbanowicz M, Ruiz-Ares G, Sepúlveda JM, et al. Urothelial cancer proteomics provides both prognostic and functional information. Sci Rep. 2017;7(1):15819. Epub 2017/11/17. doi: 10.1038/s41598-017-15920-6. PubMed PMID: 29150671; PubMed Central PMCID: PMCPMC5694001.
5. Trilla-Fuertes L, Gamez-Pozo A, M Arevalillo J, Diaz-Almiron M, Prado-Vazquez G, Zapater-Moros A, et al. Molecular characterization of breast cancer cell response to metabolic drugs. Oncotarget2018.
6. Gámez-Pozo A, Trilla-Fuertes L, Prado-Vázquez G, Chiva C, López-Vacas R, Nanni P, et al. Prediction of adjuvant chemotherapy response in triple negative breast cancer with discovery and targeted proteomics. PLoS One. 2017;12(6):e0178296. Epub 2017/06/08. doi: 10.1371/journal.pone.0178296. PubMed PMID: 28594844; PubMed Central PMCID: PMCPMC5464546.
7. Gámez-Pozo A, Ferrer NI, Ciruelos E, López-Vacas R, Martínez FG, Espinosa E, et al. Shotgun proteomics of archival triple-negative breast cancer samples. Proteomics Clin Appl. 2013;7(3-4):283-91. doi: 10.1002/prca.201200048. PubMed PMID: 23436753.

8. Cox J, Mann M. MaxQuant enables high peptide identification rates, individualized p.p.b.-range mass accuracies and proteome-wide protein quantification. Nat Biotechnol. 2008;26(12):1367-72. Epub 2008/11/26. doi: 10.1038/nbt.1511. PubMed PMID: 19029910.
9. Cox J, Neuhauser N, Michalski A, Scheltema RA, Olsen JV, Mann M. Andromeda: a peptide search engine integrated into the MaxQuant environment. J Proteome Res. 2011;10(4):1794-805. Epub 2011/01/25. doi: 10.1021/pr101065j. PubMed PMID: 21254760.
10. Deeb SJ, D'Souza RC, Cox J, Schmidt-Supprian M, Mann M. Super-SILAC allows classification of diffuse large B-cell lymphoma subtypes by their protein expression profiles. Mol Cell Proteomics. 2012;11(5):77-89. Epub 2012/03/21. doi: 10.1074/mcp.M111.015362. PubMed PMID: 22442255; PubMed Central PMCID: PMCPMC3418848.
11. Johnson WE, Li C, Rabinovic A. Adjusting batch effects in microarray expression data using empirical Bayes methods. Biostatistics. 2007;8(1):118-27. doi: 10.1093/biostatistics/kxj037. PubMed PMID: 16632515.
12. Pearl J. Probabilistic reasoning in intelligent systems: networks of plausible inference. Morgan Kaufmann2014.

13. Neapolitan RE. Learning Bayesian Networks . Series in Artificial Intelligence. Prentice Hall2004.

14. Spirtes P, Glymour C, Scheines R. Causation, Prediction, and Search. Adaptive Computation and Machine Learning. 2nd ed. The MIT Press2000.

15. Kalisch M, Maechler M, Colombo D, Maathius MH, Buehlmann P. Causal Inference Using Graphical Models with the R Package pcalg. Journal of Statistical Software2012. p. 1-26.

16. Kalisch M, Bühlmann P. Estimating high-dimensional directed acyclic graphs with the PC algorithm. Journal of Machine Learning Research. 2007;8:613-36.
17. Kalisch M, Bülhlmann P. Estimating high-dimensional directed acyclic graphs with the PC algorithm. Journal of Computational and Graphial Statistics. 2008;17(4):613-36.
18. Harris N, M D. PC algorithm for nonparanormal graphical models. Journal of Machine Learning Research. 2013;14:3365-83.
19. R Core Team. R: A language and environment for statistical computing. Vienna,Austria. R Foundation for Stattistical Computing, 2013.
20. Gentleman R, Whalen E, Huber W, Falcon S. graph: A package to handle graph data structures. R package version 1.54.0.

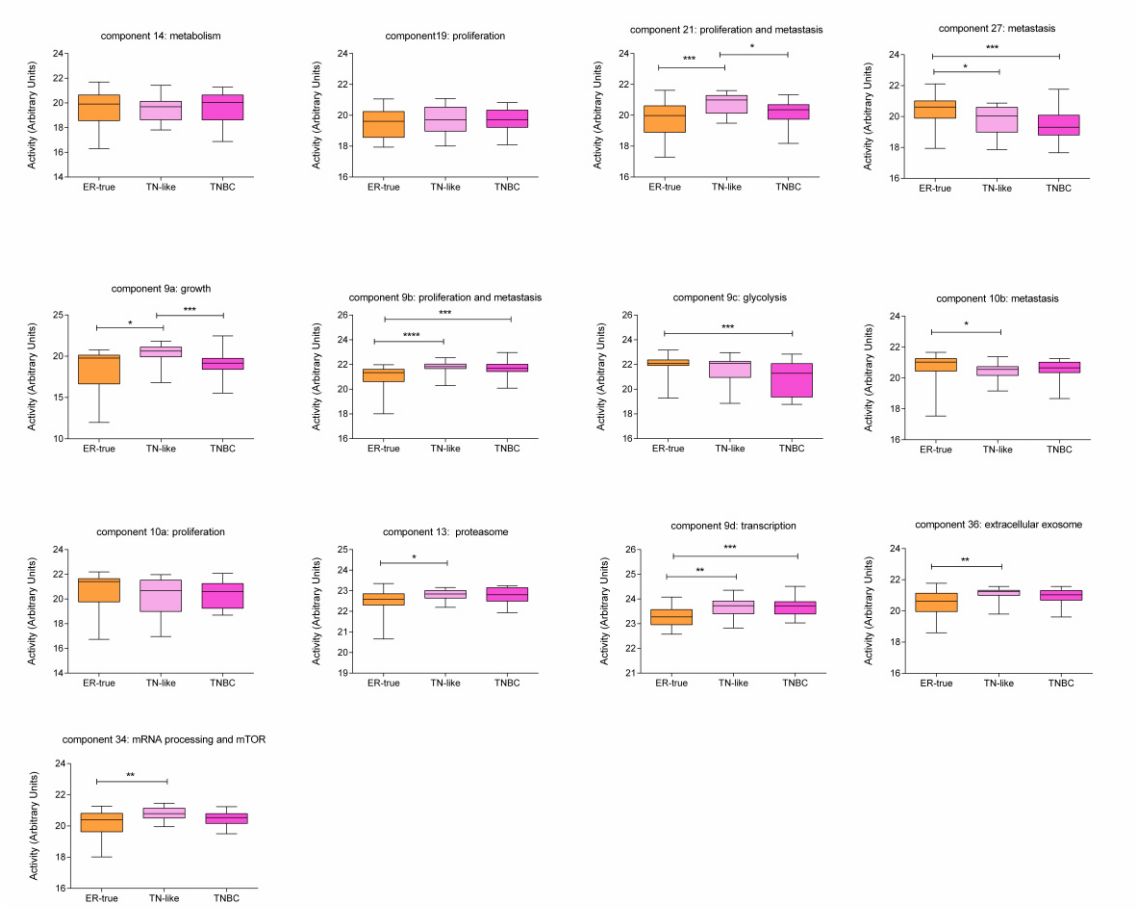
21. Dannenfelser R, Clark NR, Ma'ayan A. Genes2FANs: connecting genes through functional association networks. BMC Bioinformatics. 2012;13:156. doi: 10.1186/1471-2105-13-156. PubMed PMID: 22748121; PubMed Central PMCID: PMCPMC3472228.
22. Huang dW, Sherman BT, Lempicki RA. Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. Nat Protoc. 2009;4(1):44-57. doi: 10.1038/nprot.2008.211. PubMed PMID: 19131956.
23. Simon R. Roadmap for developing and validating therapeutically relevant genomic classifiers. J Clin Oncol. 2005;23(29):7332-41. Epub 2005/09/06. doi: 10.1200/JCO.2005.02.8712. PubMed PMID: 16145063.
24. Shannon P, Markiel A, Ozier O, Baliga NS, Wang JT, Ramage D, et al. Cytoscape: a software environment for integrated models of biomolecular interaction networks. Genome Res. 2003;13(11):2498-504. doi: 10.1101/gr.1239303. PubMed PMID: 14597658; PubMed Central PMCID: PMCPMC403769.
25. Kim DG, Choi JW, Lee JY, Kim H, Oh YS, Lee JW, et al. Interaction of two translational components, lysyl-tRNA synthetase and p40/37LRP, in plasma membrane promotes laminin-

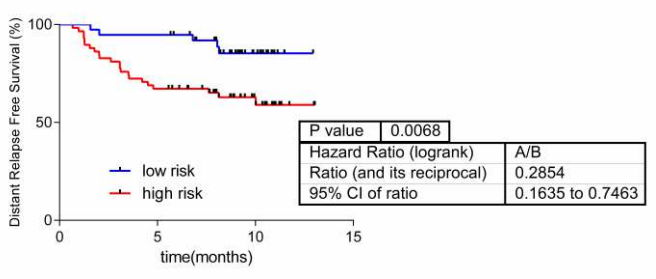
418 dependent cell migration. FASEB J. 2012;26(10):4142-59. Epub 2012/07/02. doi: 10.1096/fj.12-
419 207639. PubMed PMID: 22751010.
420 26. Jin Q, Pulipati NR, Zhou W, Staub CM, Liotta LA, Mulder KM. Role of km23-1 in
421 RhoA/actin-based cell migration. Biochem Biophys Res Commun. 2012;428(3):333-8. Epub
422 2012/10/15. doi: 10.1016/j.bbrc.2012.10.047. PubMed PMID: 23079622; PubMed Central
423 PMCID: PMCPMC3513371.
424 27. Madak-Erdogan Z, Ventrella R, Petry L, Katzenellenbogen BS. Novel roles for ERK5 and
425 cofilin as critical mediators linking ERα-driven transcription, actin reorganization, and
426 invasiveness in breast cancer. Mol Cancer Res. 2014;12(5):714-27. Epub 2014/02/06. doi:
427 10.1158/1541-7786.MCR-13-0588. PubMed PMID: 24505128; PubMed Central PMCID:
428 PMCPMC4020978.
429 28. Katoh H, Hiramoto K, Negishi M. Activation of Rac1 by RhoG regulates cell migration. J
430 Cell Sci. 2006;119(Pt 1):56-65. Epub 2005/12/08. doi: 10.1242/jcs.02720. PubMed PMID:
431 16339170.
432 29. Järvinen TA, Prince S. Decorin: A Growth Factor Antagonist for Tumor Growth
433 Inhibition. Biomed Res Int. 2015;2015:654765. Epub 2015/11/30. doi: 10.1155/2015/654765.
434 PubMed PMID: 26697491; PubMed Central PMCID: PMCPMC4677162.
435 30. Mohan J, Morén B, Larsson E, Holst MR, Lundmark R. Cavin3 interacts with cavin1 and
436 caveolin1 to increase surface dynamics of caveolae. J Cell Sci. 2015;128(5):979-91. Epub
437 2015/01/14. doi: 10.1242/jcs.161463. PubMed PMID: 25588833.
438 31. Jin Q, Ding W, Mulder KM. The TGFβ receptor-interacting protein km23-1/DYNLRB1
439 plays an adaptor role in TGFβ1 autoinduction via its association with Ras. Journal of Biological
440 Chemistry. 2012;287(31):26453-63.
441 32. Moustakas A, Heldin C-H. Dynamic control of TGF-β signaling and its links to the
442 cytoskeleton. FEBS letters. 2008;582(14):2051-65.
443 33. Green DR, Ferguson T, Zitvogel L, Kroemer G. Immunogenic and tolerogenic cell death.
444 Nature reviews Immunology. 2009;9(5):353.
445 34. Castillo J, Bernard V, San Lucas FA, Allenson K, Capello M, Kim DU, et al. Surfaceome
446 profiling enables isolation of cancer-specific exosomal cargo in liquid biopsies from pancreatic
447 cancer patients. Ann Oncol. 2017. Epub 2017/09/25. doi: 10.1093/annonc/mdx542. PubMed
448 PMID: 29045505.
449 35. Cho HY, Ul Mushtaq A, Lee JY, Kim DG, Seok MS, Jang M, et al. Characterization of the
450 interaction between lysyl-tRNA synthetase and laminin receptor by NMR. FEBS letters.
451 2014;588(17):2851-8.
452 36. Hiramoto K, Negishi M, Katoh H. Dock4 is regulated by RhoG and promotes Rac-
453 dependent cell migration. Experimental cell research. 2006;312(20):4205-16.
454 37. Feugaing DDS, Götte M, Viola M. More than matrix: the multifaceted role of decorin in
455 cancer. European journal of cell biology. 2013;92(1):1-11.
456 38. Cawthorn TR, Moreno JC, Dharsee M, Tran-Thanh D, Ackloo S, Zhu PH, et al. Proteomic
457 analyses reveal high expression of decorin and endoplasmin (HSP90B1) are associated with
458 breast cancer metastasis and decreased survival. PloS one. 2012;7(2):e30992.
459 39. Nam SH, Kang M, Ryu J, Kim HJ, Kim D, Kim DG, et al. Suppression of lysyl-tRNA
460 synthetase, KRS, causes incomplete epithelial-mesenchymal transition and ineffective cell-
461 extracellular matrix adhesion for migration. Int J Oncol. 2016;48(4):1553-60. Epub 2016/02/08.
462 doi: 10.3892/ijo.2016.3381. PubMed PMID: 26891990.

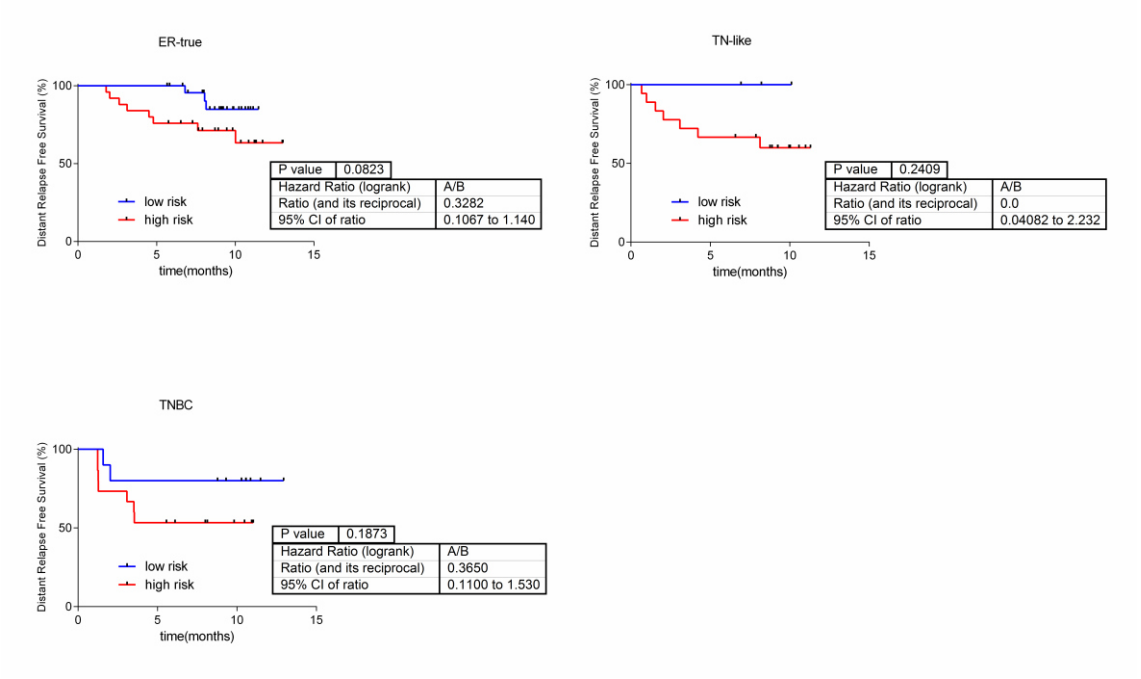
Supporting information captions

S1 Table: Clinical characteristics of the patient cohort.

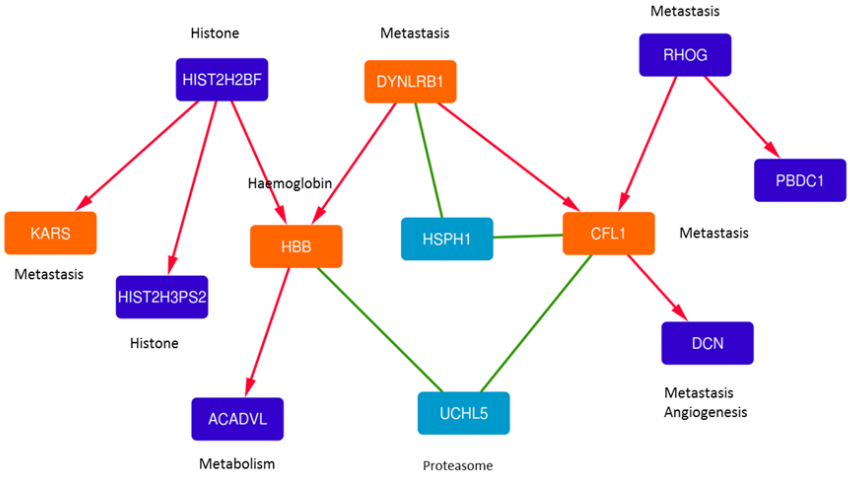
S1 File: Components defined by DAG analysis.







468



469